

An aerial view of a city at sunset, with a large orange and purple sky. Overlaid on the image are several glowing white lines that form a network or orbital pattern, connecting various points across the city and sky. The right side of the image is partially obscured by a semi-transparent orange shape.

# Accelerating Software 2.0

Foundations for Next-Generation  
Computer Systems

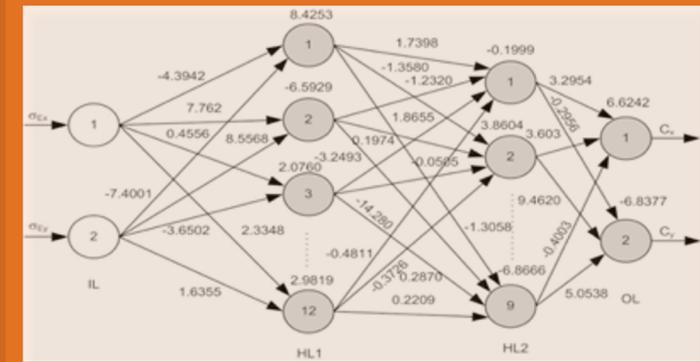
**Christopher Aberger**

Director of Software Engineering



# Software 1.0 vs Software 2.0

```
37 #include <iostream>
38 using namespace std;
39
40 int _tmain (int argc, _TCHAR* argv[])
41 {
42
43     int iVal1 = 0, iVal2 = 0, iVal3 = 0;
44
45     printf("Enter three numbers:");
46     scanf("%d %d %d", &iVal1, &iVal2, &iVal3);
47
48     if (iVal1 >= iVal2)
49     {
50         if(iVal1 >= iVal3)
51             printf("Largest number = %.2d", iVal1);
52         else
53             printf("Largest number = %.2d", iVal3);
54     }
55     else
56     {
57         if(iVal2 >= iVal3)
58             printf("Largest number = %.2d", iVal2);
59         else
60             printf("Largest number = %.2d", iVal3);
61     }
62
63     getchar ();
64     return 0;
65 }
```

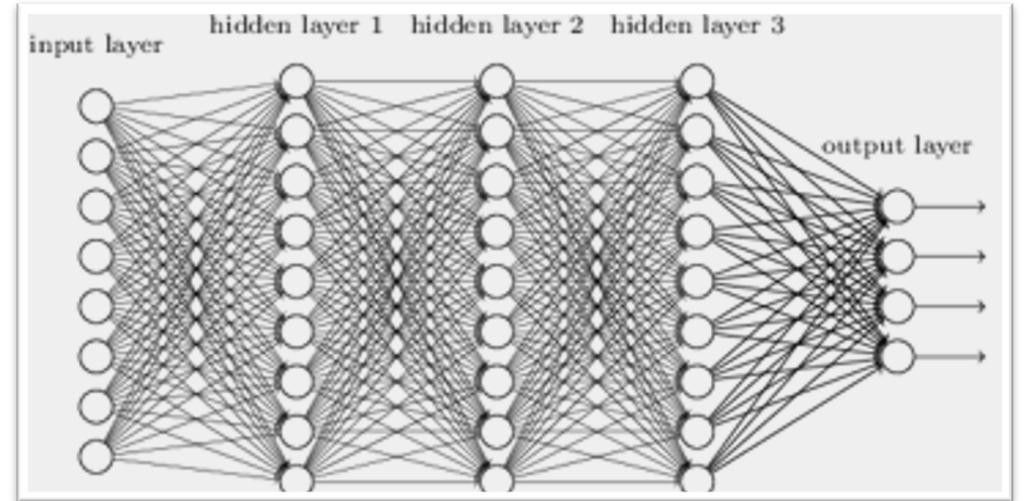


- Written in code (C++, ...)
- Requires domain expertise
  - Decompose the problem
  - Design algorithms
  - Compose into a system

- Programmer input: training data
- Written in the weights of a neural network model by optimization
- Reduced lines of code

Andrej Karpathy. Scaled ML 2018 talk

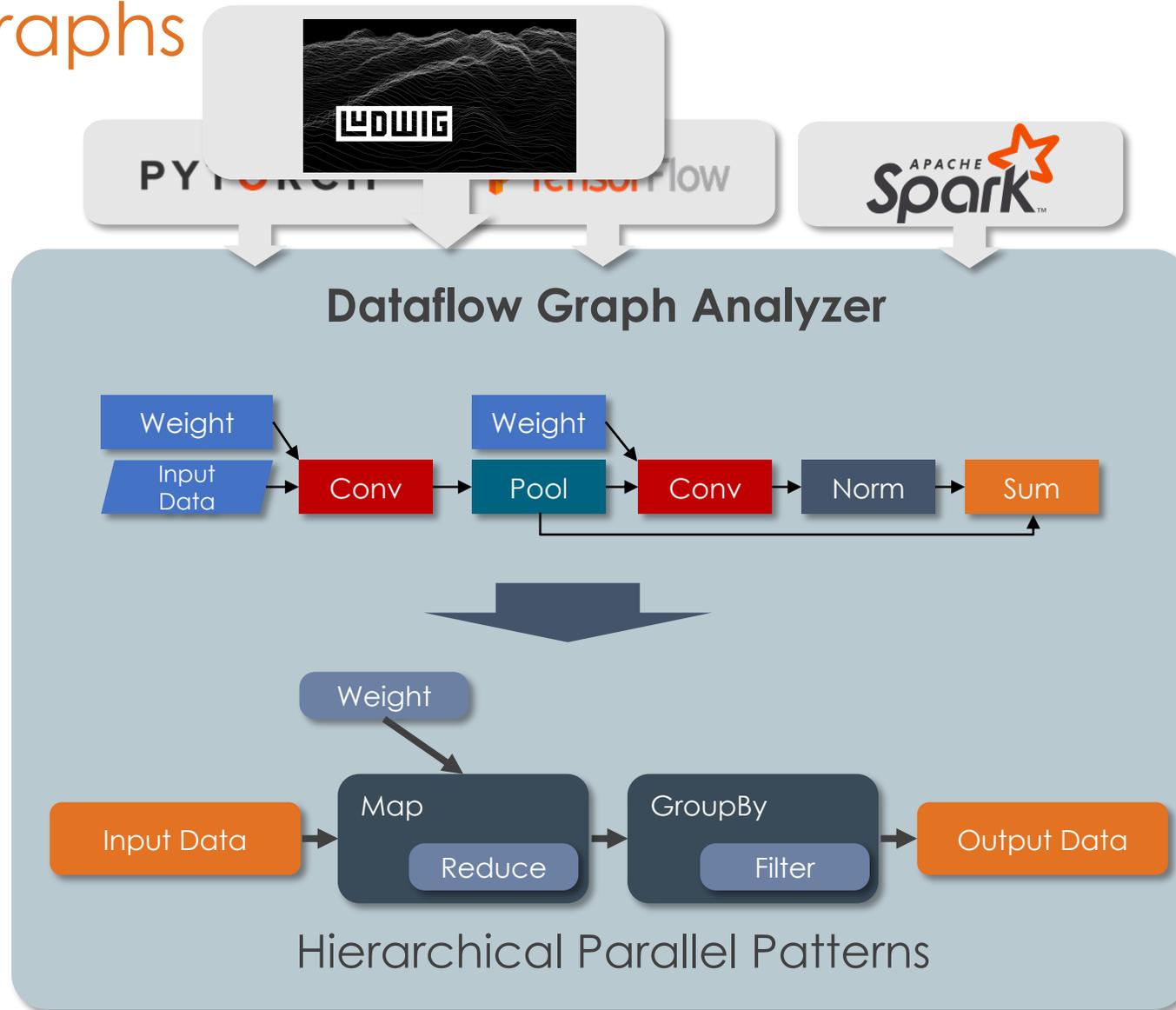
# Software 2.0 is Dataflow



## 1000x Productivity

Google shrinks language translation code from 500k imperative LoC to **500 lines of dataflow (TensorFlow)**

# Dataflow Graphs



# Software 2.0 is replacing Software 1.0

## The Case for Learned Index Structures

Tim Kraska\*

MIT

Cambridge, MA

kraska@mit.edu

Alex Beutel

Ed H. Chi

## HoloClean: Holistic Data Repairs with Probabilistic Inference

Jeffrey Dean

Google Inc

## Snorkel: Rapid Training Data Creation with Weak Supervision

Alexander Ratner   Stephen H. Bach   Henry Ehrenberg

Jason Fries   Sen Wu   Christopher Ré

Stanford University  
Stanford, CA, USA

{ajratner, bach, henryre, jfries, senwu, chrismre}@cs.stanford.edu

Christopher Ré\*  
Waterloo

## AI FOR SCIENCE

**RICK STEVENS**  
**VALERIE TAYLOR**

*Argonne National Laboratory*  
July 22-23, 2019

**JEFF NICHOLS**  
**ARTHUR BARNEY MACCABE**

*Oak Ridge National Laboratory*  
August 21-23, 2019

**KATHERINE YELICK**  
**DAVID BROWN**

*Lawrence Berkeley*  
*National Laboratory*  
September 11-12, 2019

# Next gen Software 2.0 systems need support for



## **Hierarchical parallel pattern Dataflow**

Natural ML execution model



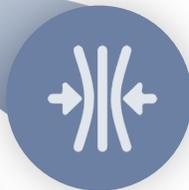
## **Terabyte sized models**

Higher accuracy



## **Sparsity**

Graph based neural networks



## **Flexible mapping**

Model and data parallelism



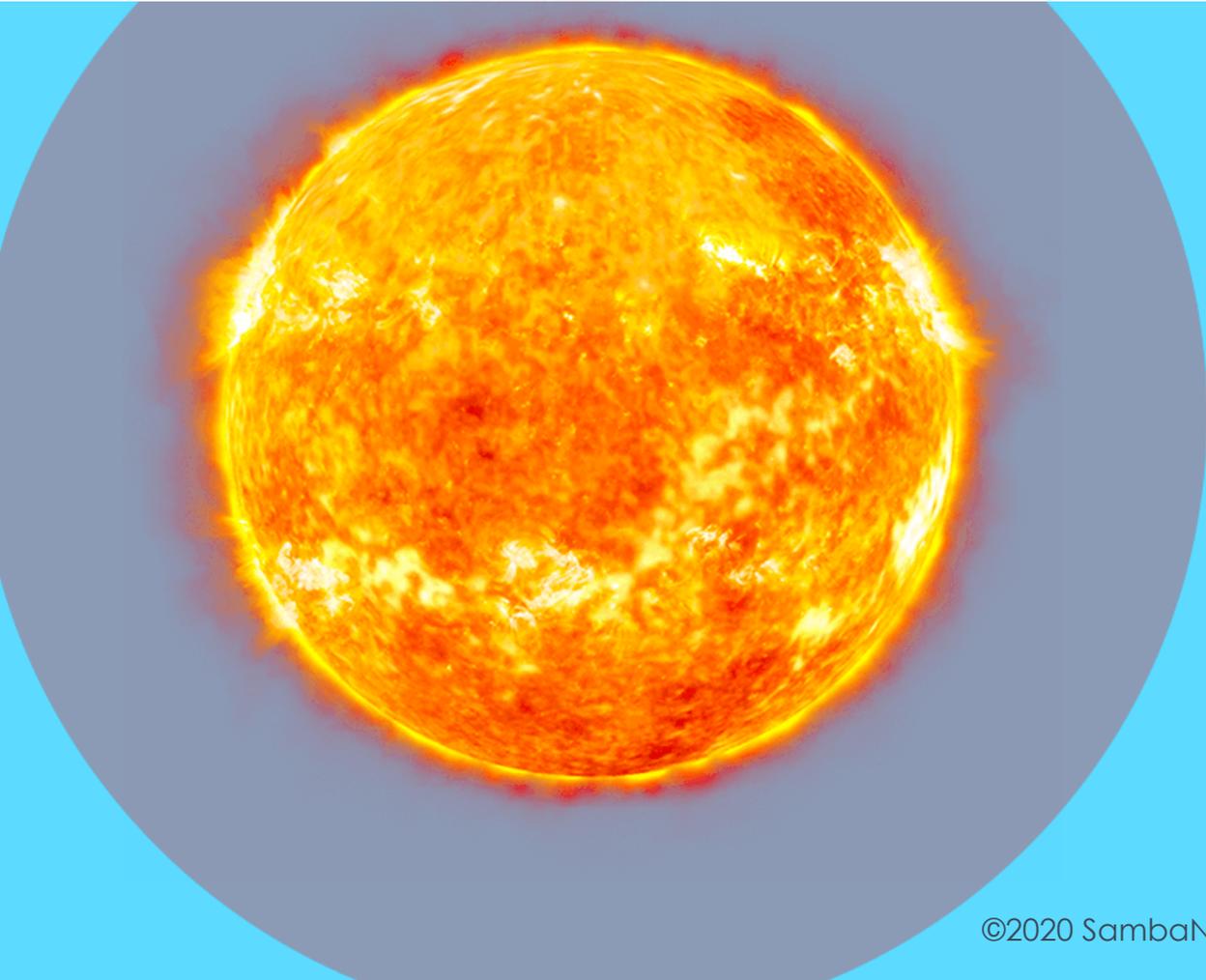
## **Data processing**

SQL in inner loop of ML training

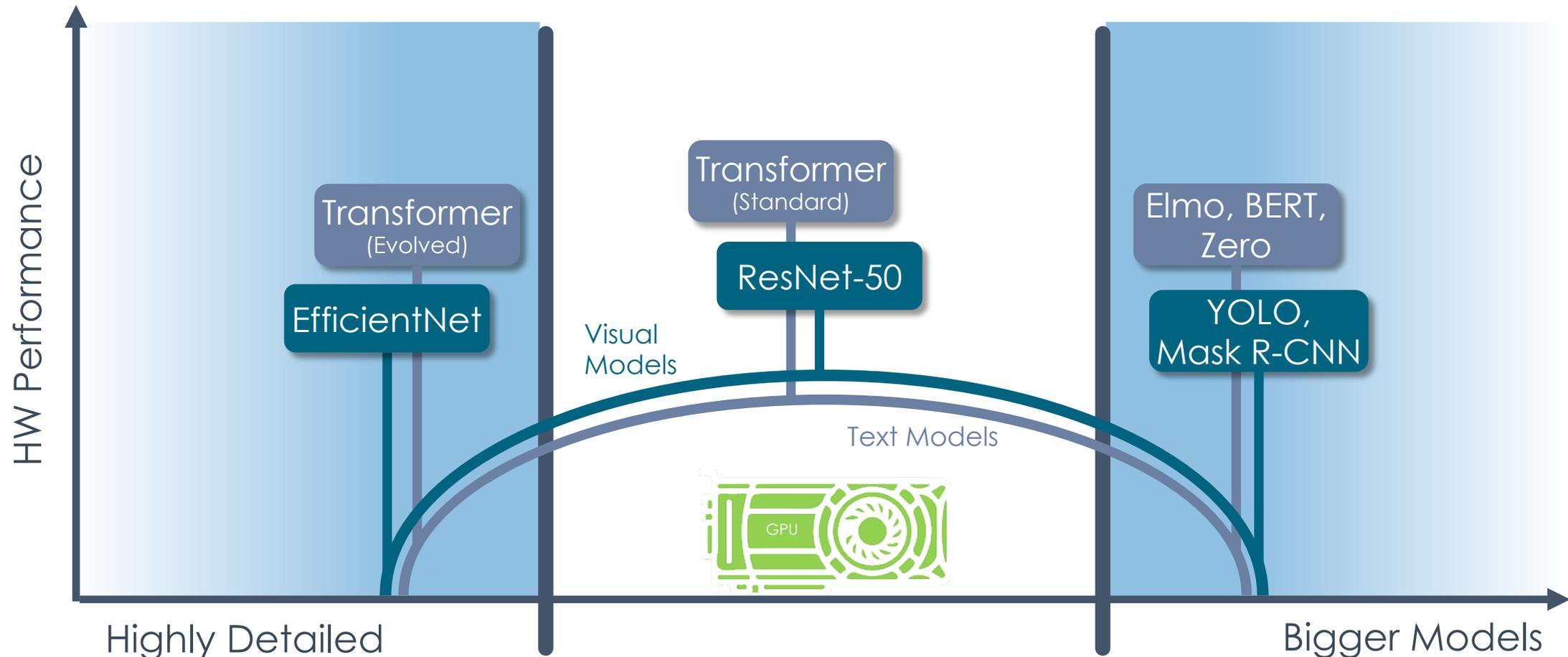
Too Hot

Goldilocks  
Zone

Too Cold



# Yesterday's Goldilocks Zone is Constraining Progress



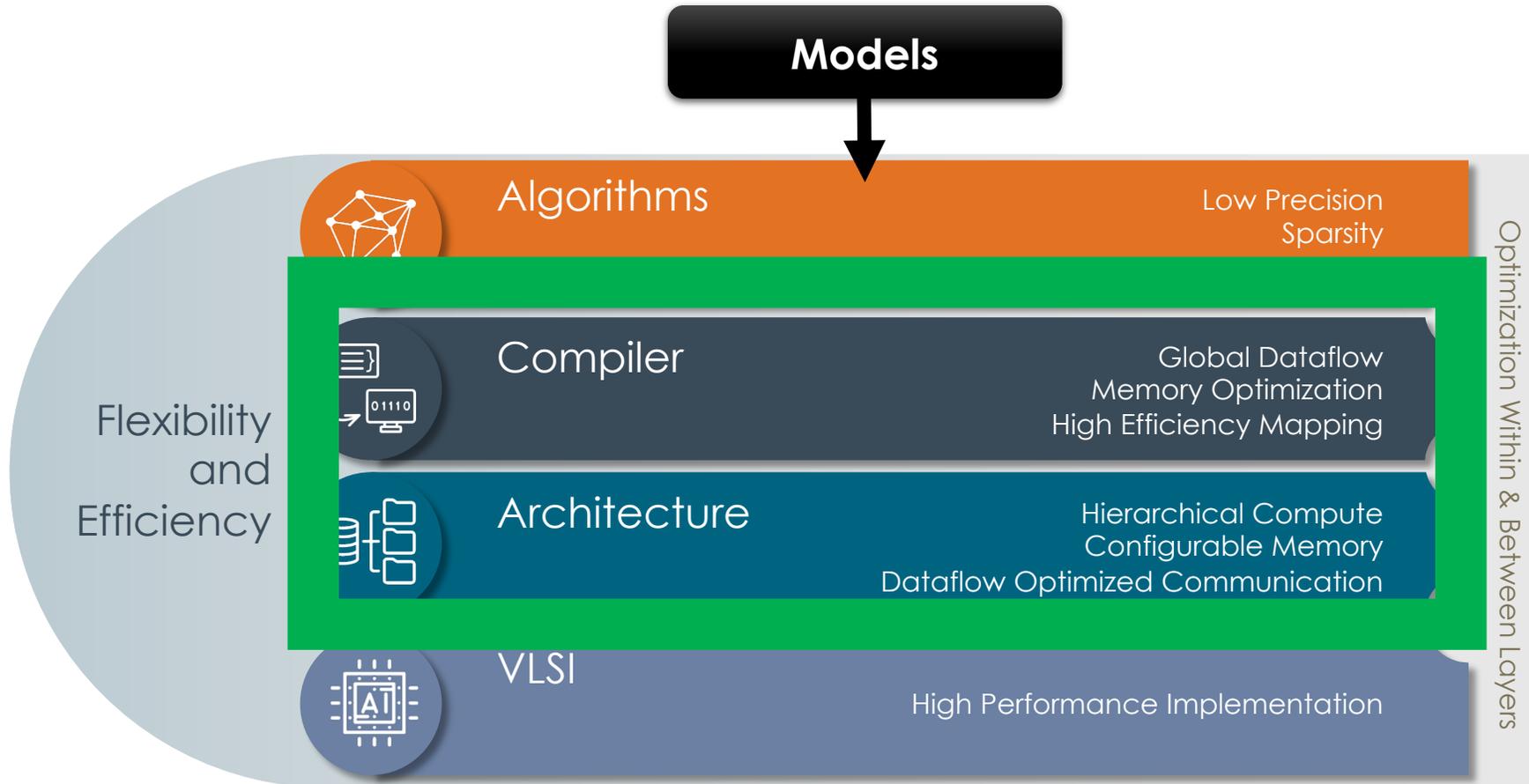
Faster, Higher Quality on Today's models—better on Tomorrow's models

# How do we break out of the Goldilocks Zone?

*Fundamental advances required at all layers of the stack.*

# The SambaNova Systems Advantage: Reconfigurable Dataflow Architecture

Full stack co-engineering yields optimizations where best delivered with the highest impact



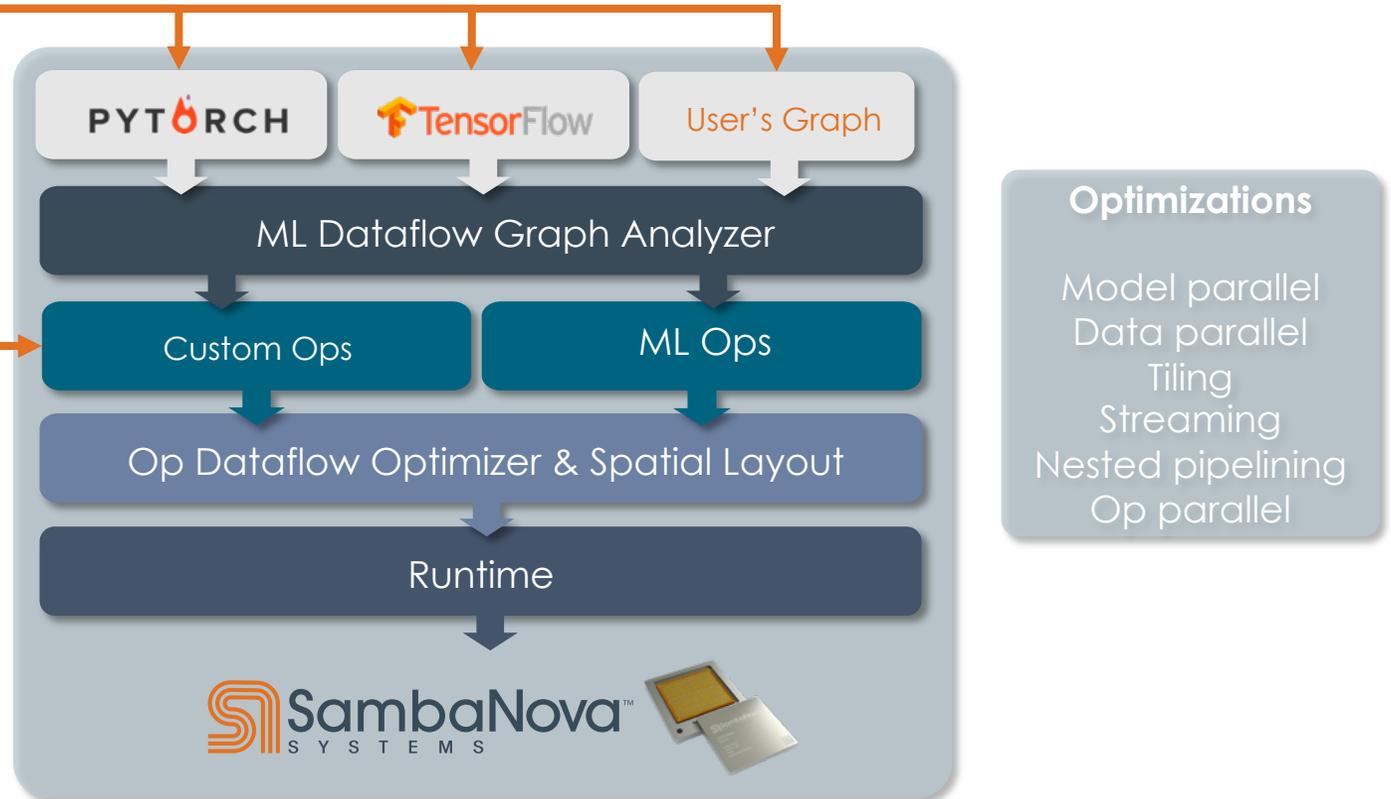
# SambaFlow Open Software for DataScale Systems

## Graph Entry Points

- Write to OSS ML frameworks or user's graph
- Push-button automation path

## API Entry Point

- User programs to DSL
- Mix of manual and automatic



# SambaNova Systems Cardinal SN10 RDU



## The Chip

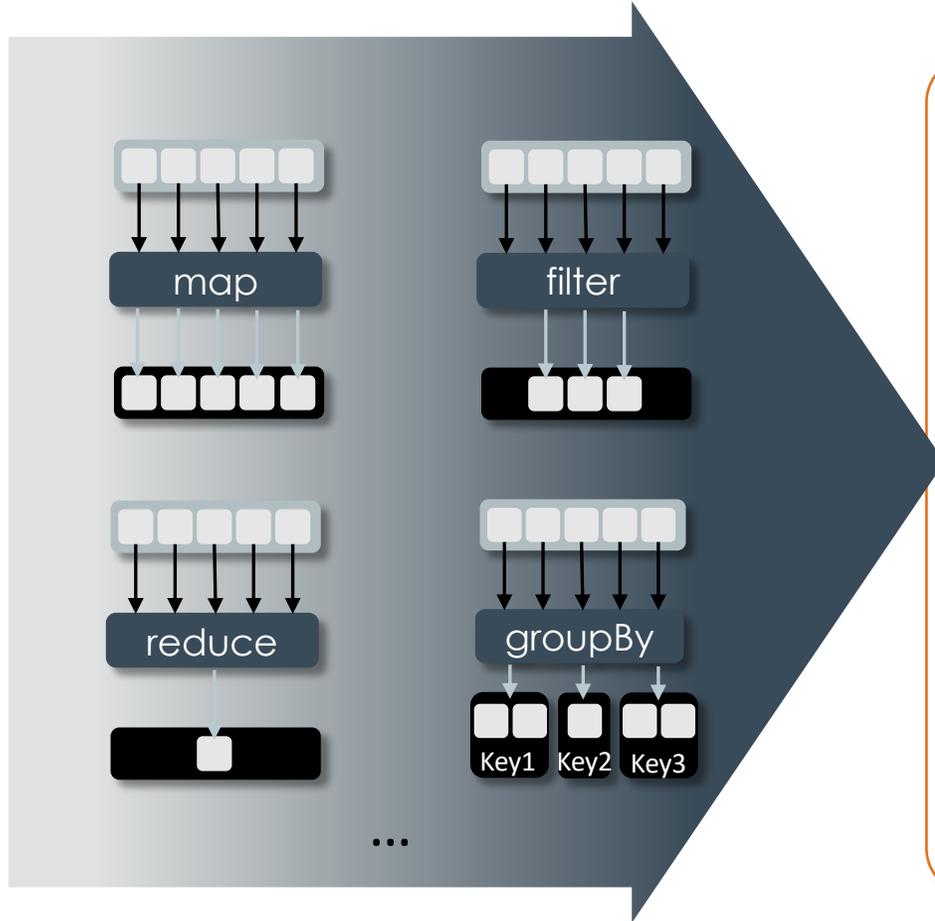
- First Reconfigurable Dataflow Unit (RDU)
- TSMC 7nm
- 40B transistors
- 50 Km of wire
- 100s of TFLOPS
- 100s MB on chip
- Direct interfaces to TBs off chip

## The System

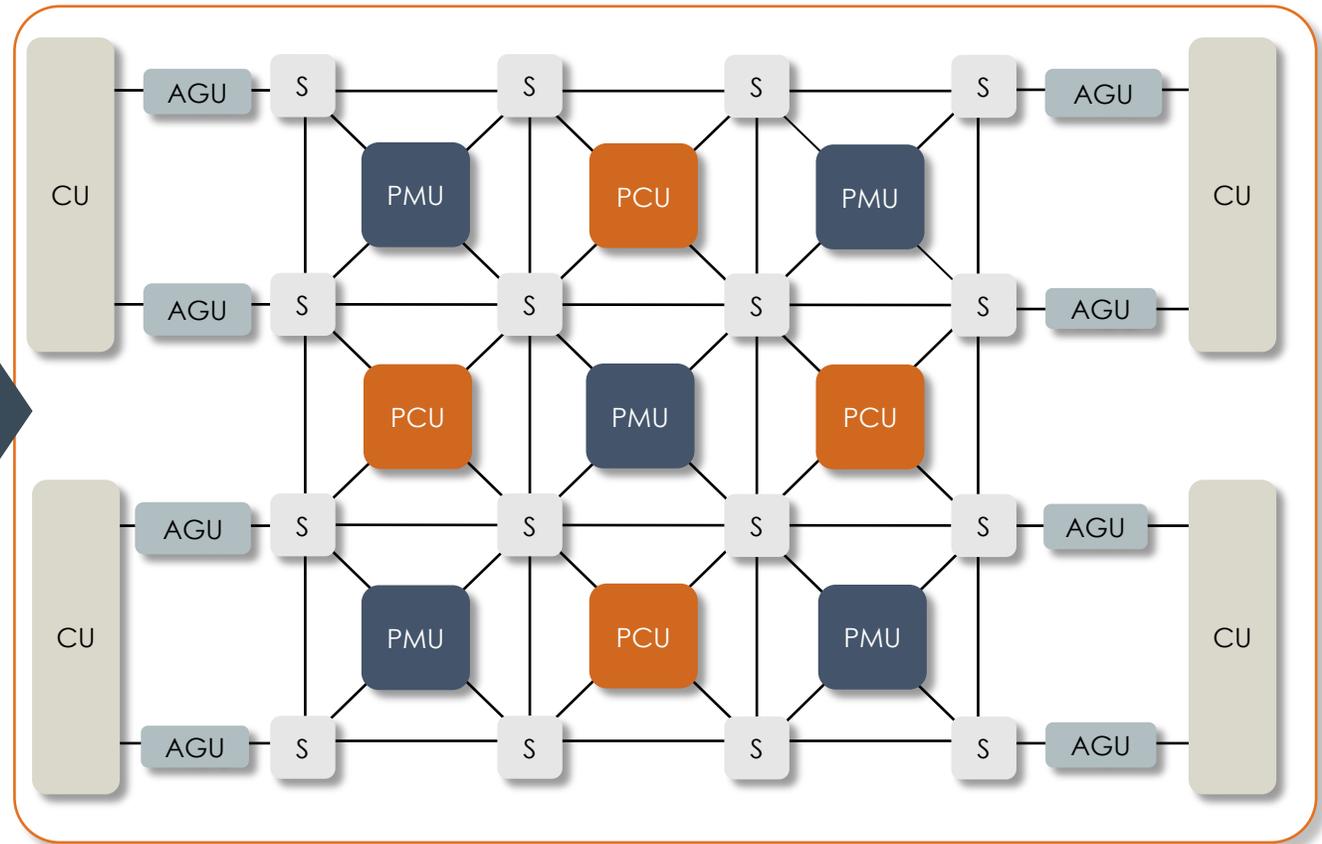
Open standard rack,  
Open standard form factor,  
Open standard power,  
Open standard cooling,  
Open standard operations ...

# Reconfigurable Dataflow Unit (RDU)

Parallel Patterns

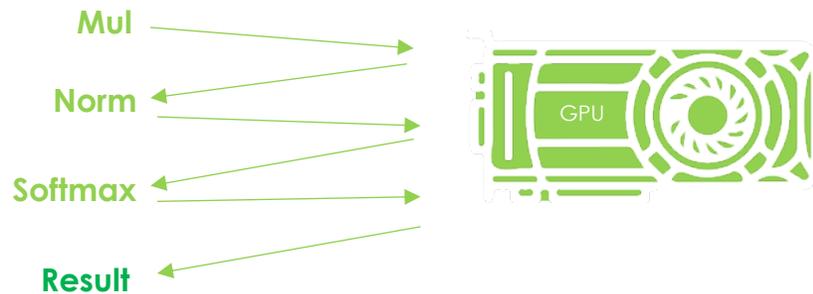


Array of reconfigurable compute, memory and communication

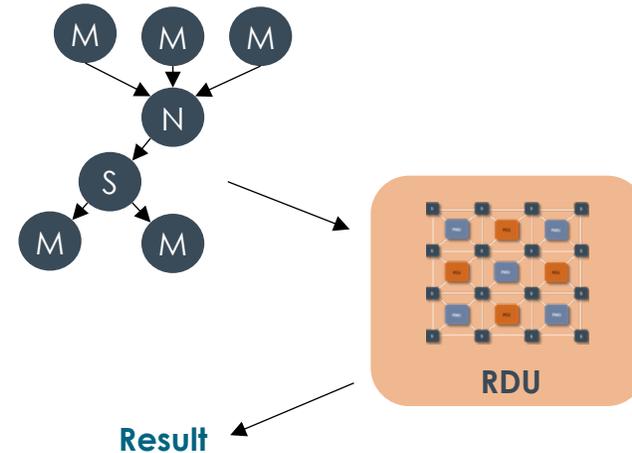


# Spatial Dataflow Within an RDU

The old way:  
kernel-by-kernel

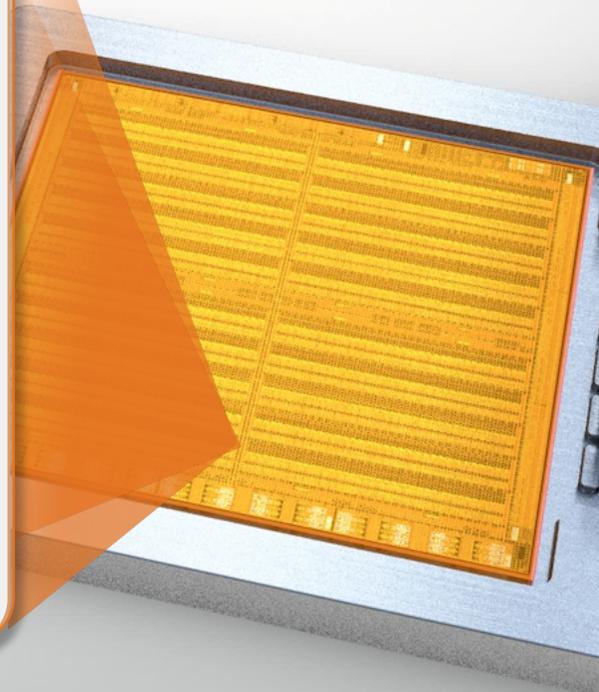
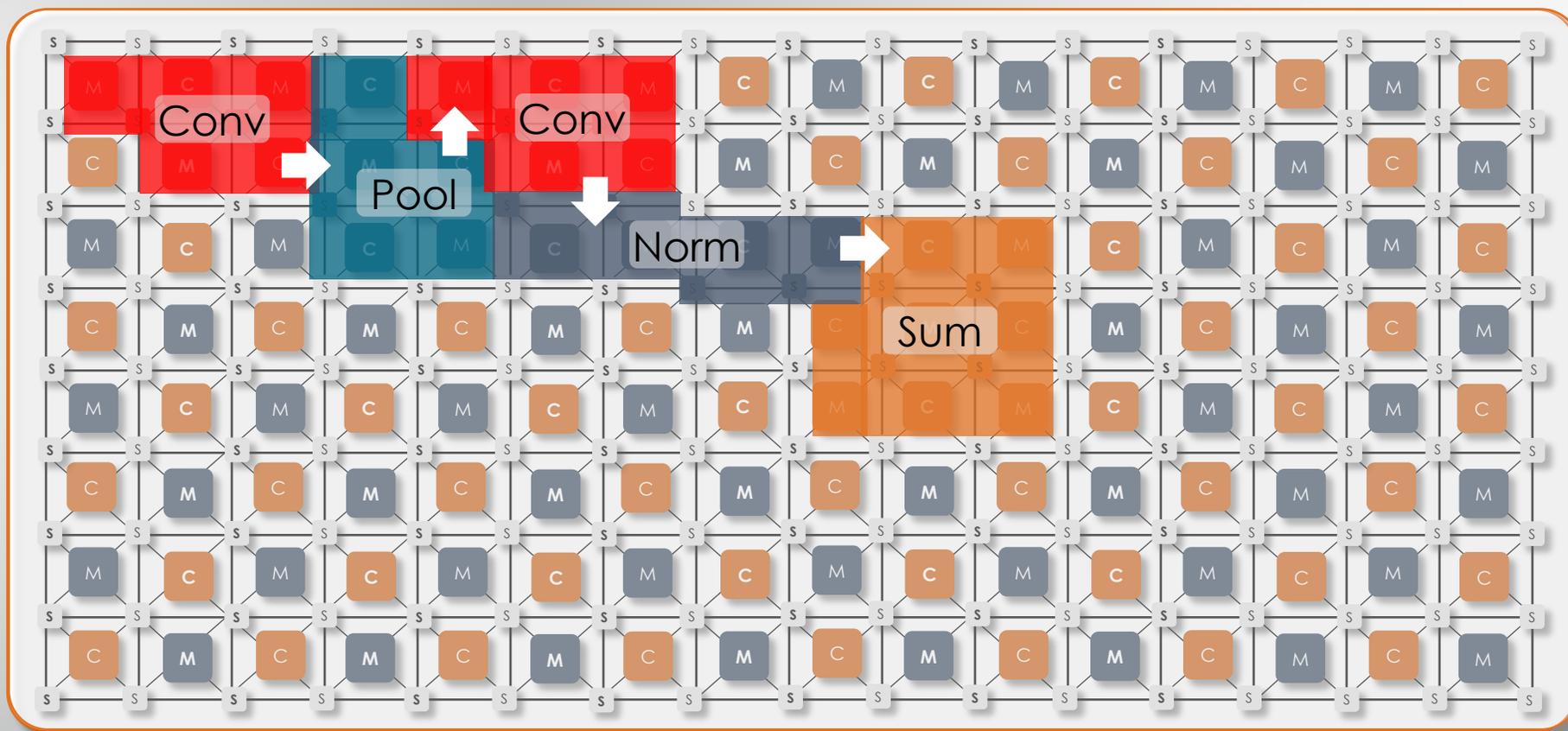
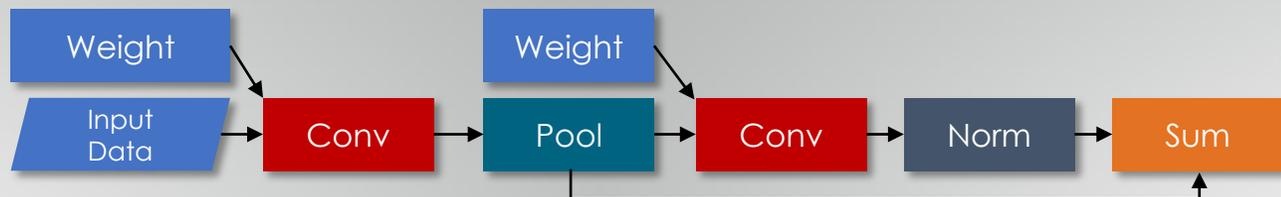


The Dataflow way:  
spatial

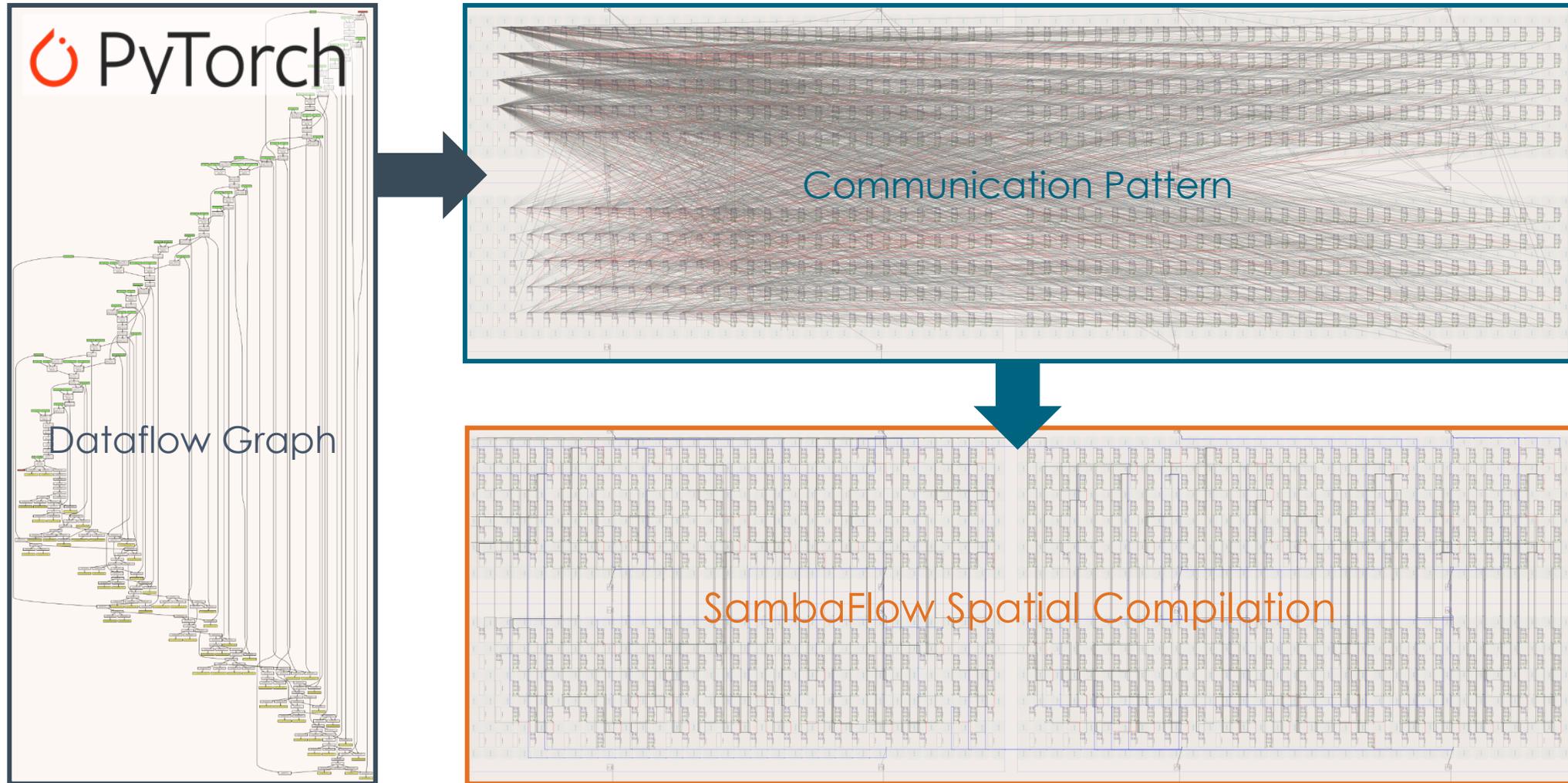


SambaFlow eliminates overhead and  
maximizes utilization

# Rapid Dataflow Compilation to RDU

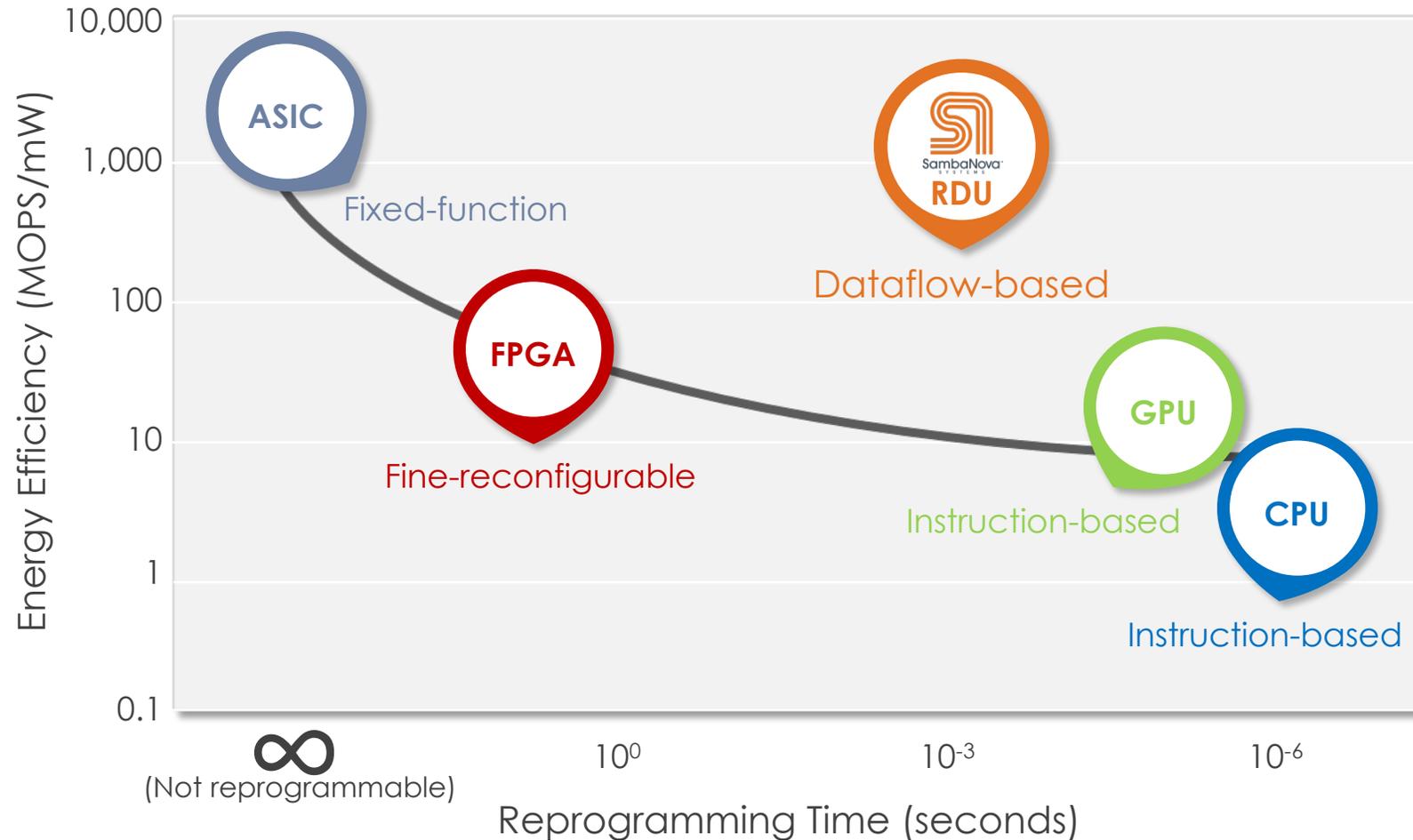


# SambaFlow Produces Highly Optimized Spatial Mappings



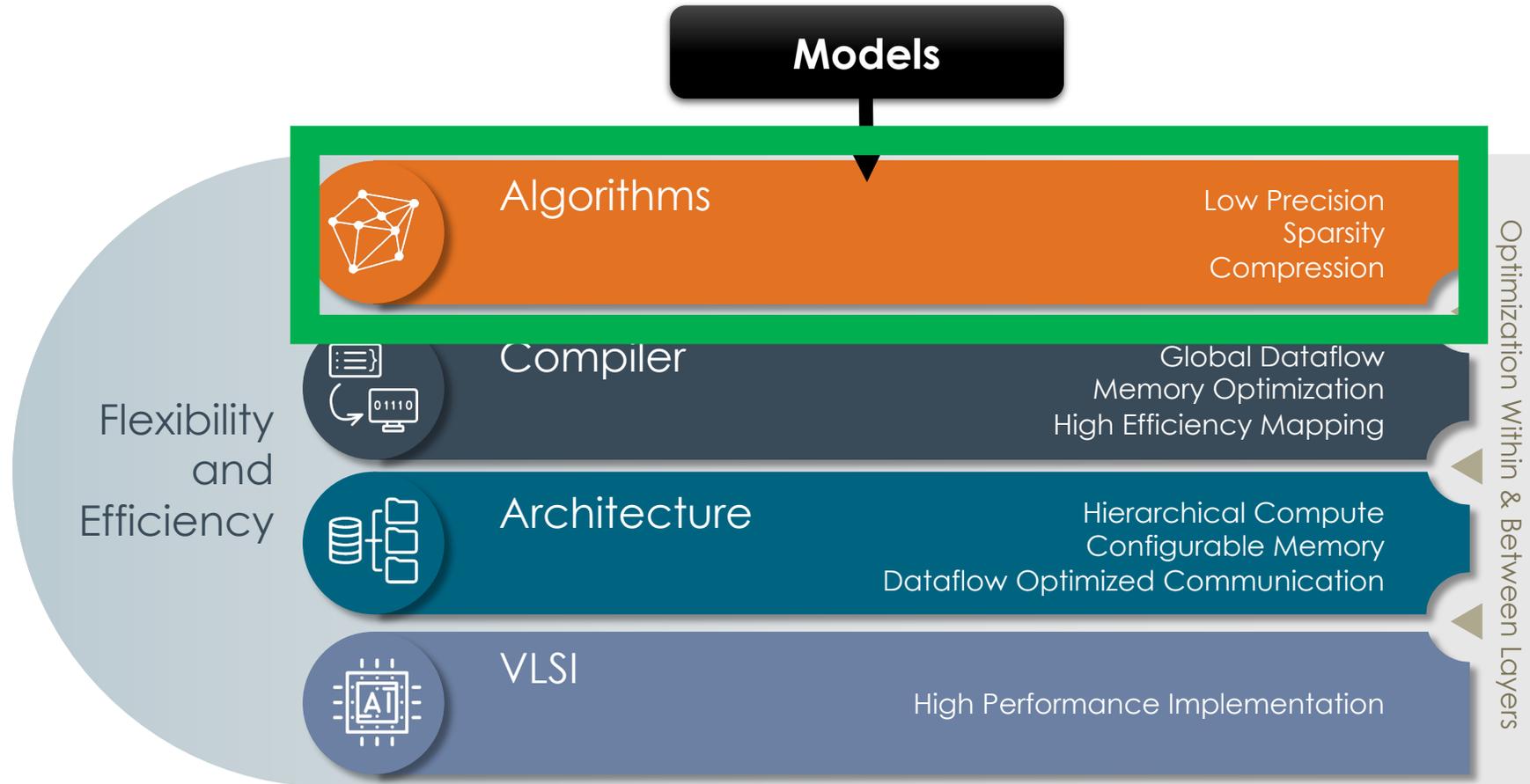
# Uncompromised Programmability and Efficiency

Breaking out of the programmability vs. efficiency tradeoff curve



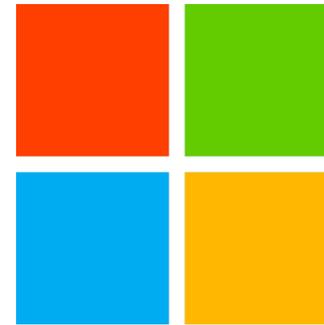
# The SambaNova Systems Advantage: Reconfigurable Dataflow Architecture

Full stack co-engineering yields optimizations where best delivered with the highest impact

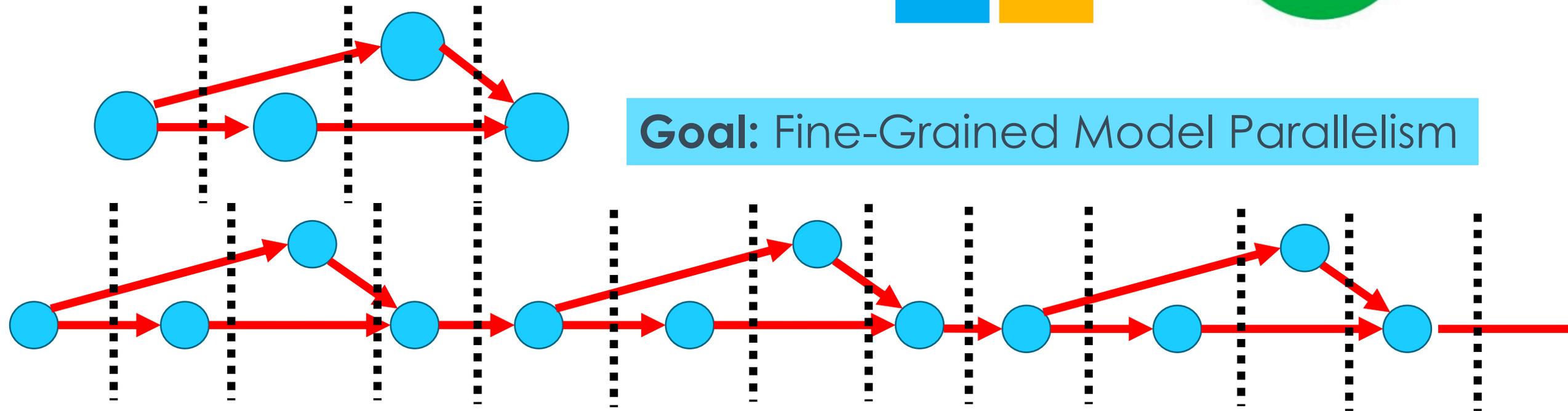


# Model (Pipeline) Parallelism: Are we there yet?

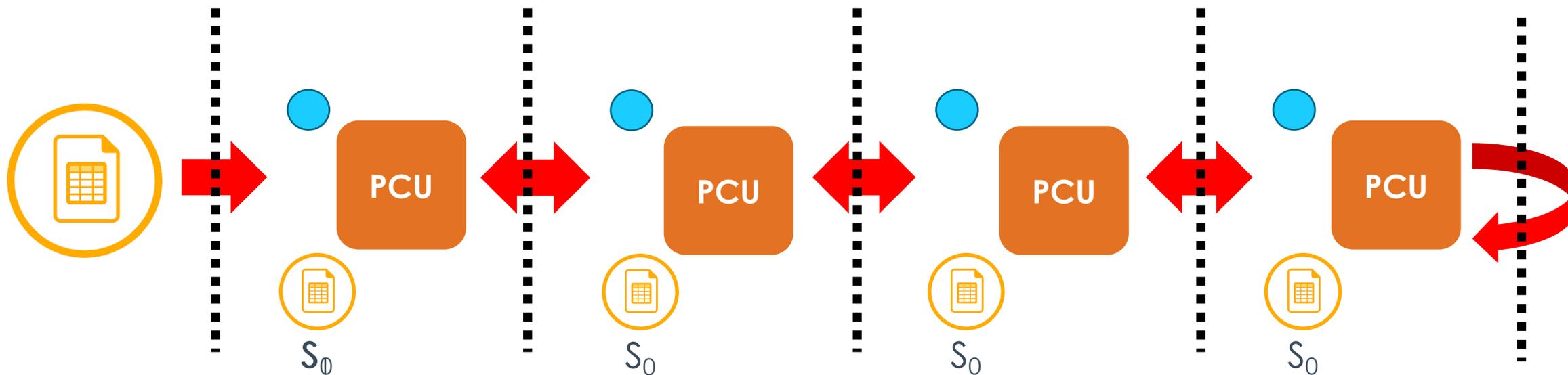
- 1. Course Grained
- 2. HW Cost



**Goal:** Fine-Grained Model Parallelism

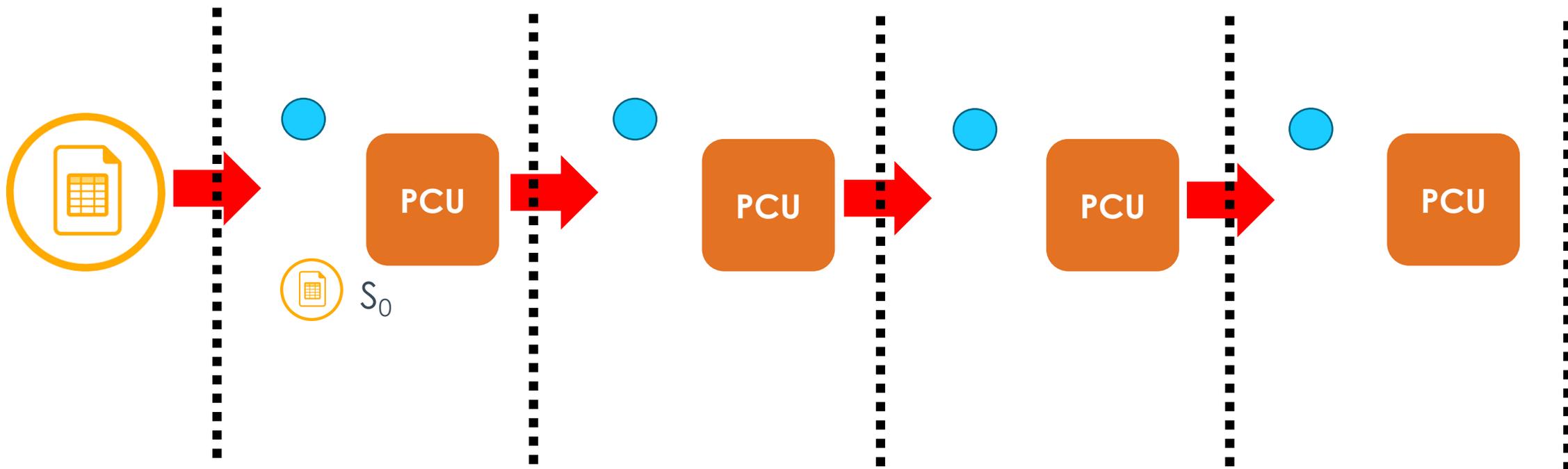


# HW Cost: GPipe

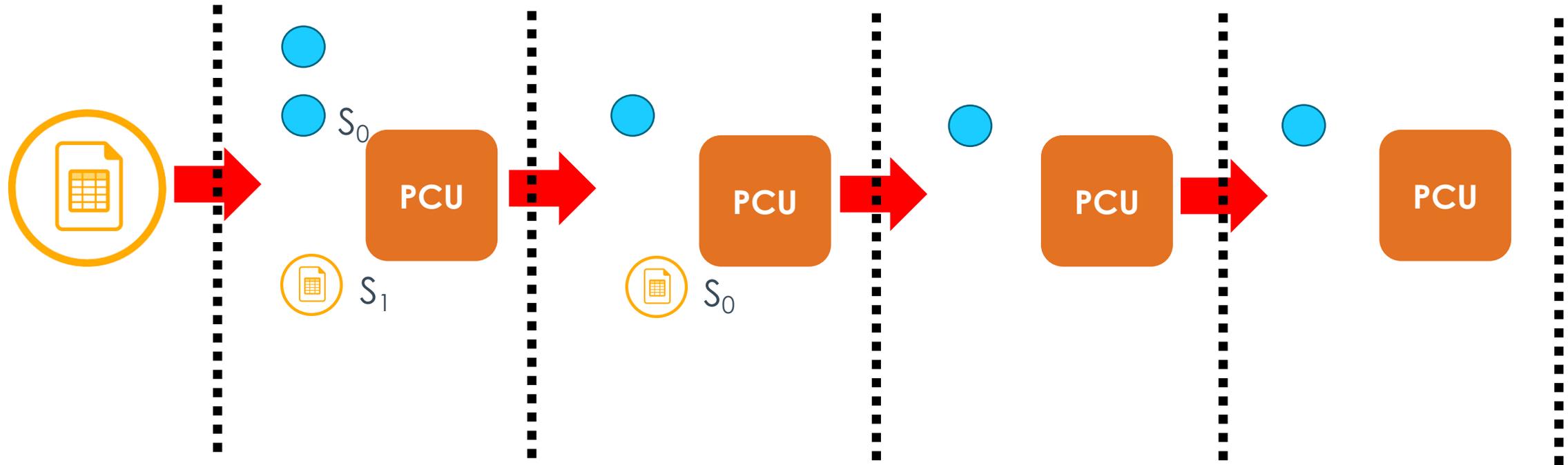


**Panic:** Sacrifices latency for synchronous execution!

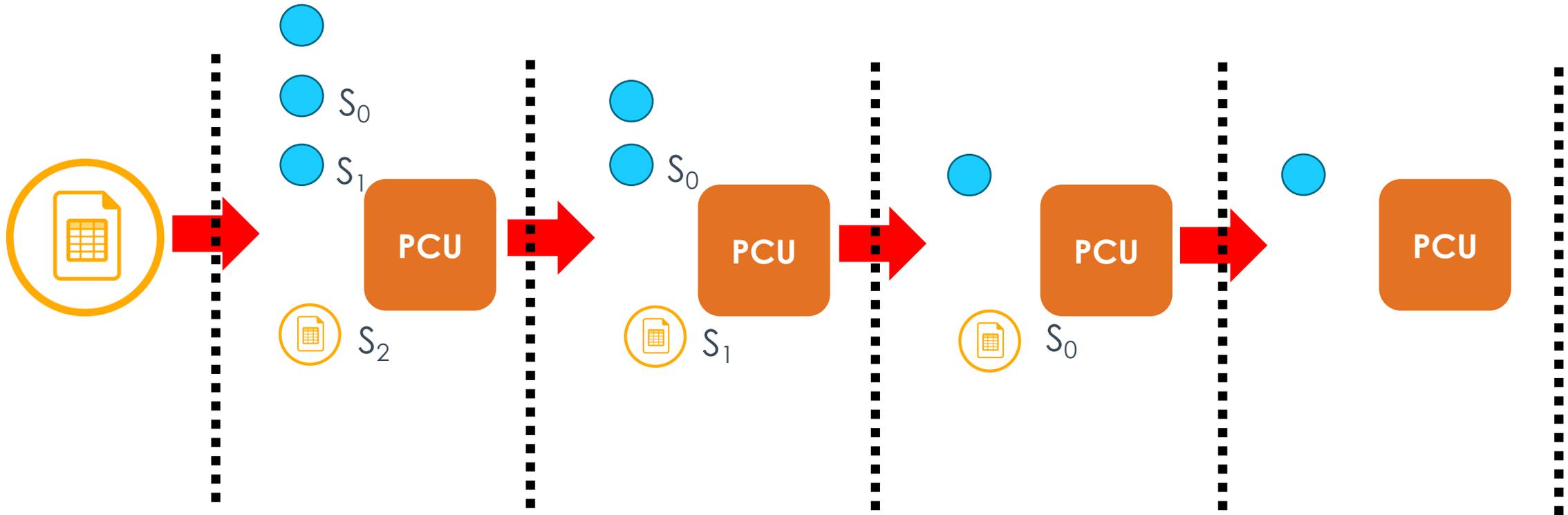
# HW Cost: PipeDream



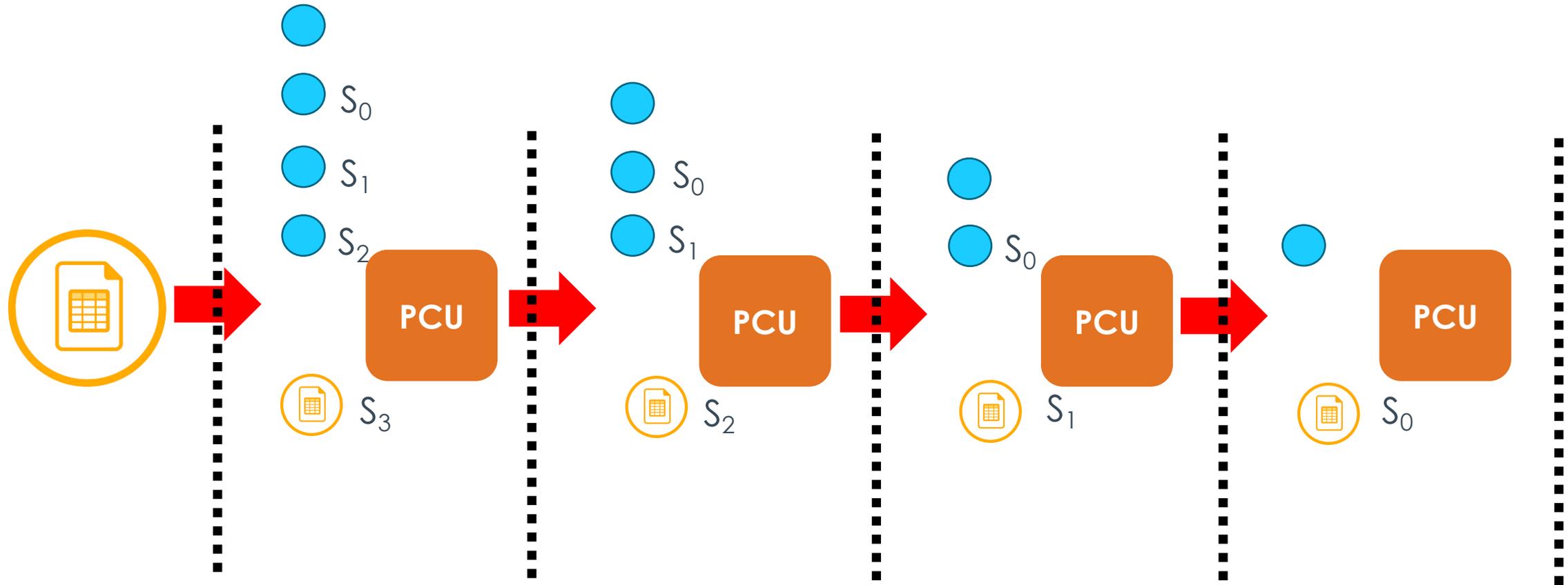
# HW Cost: PipeDream



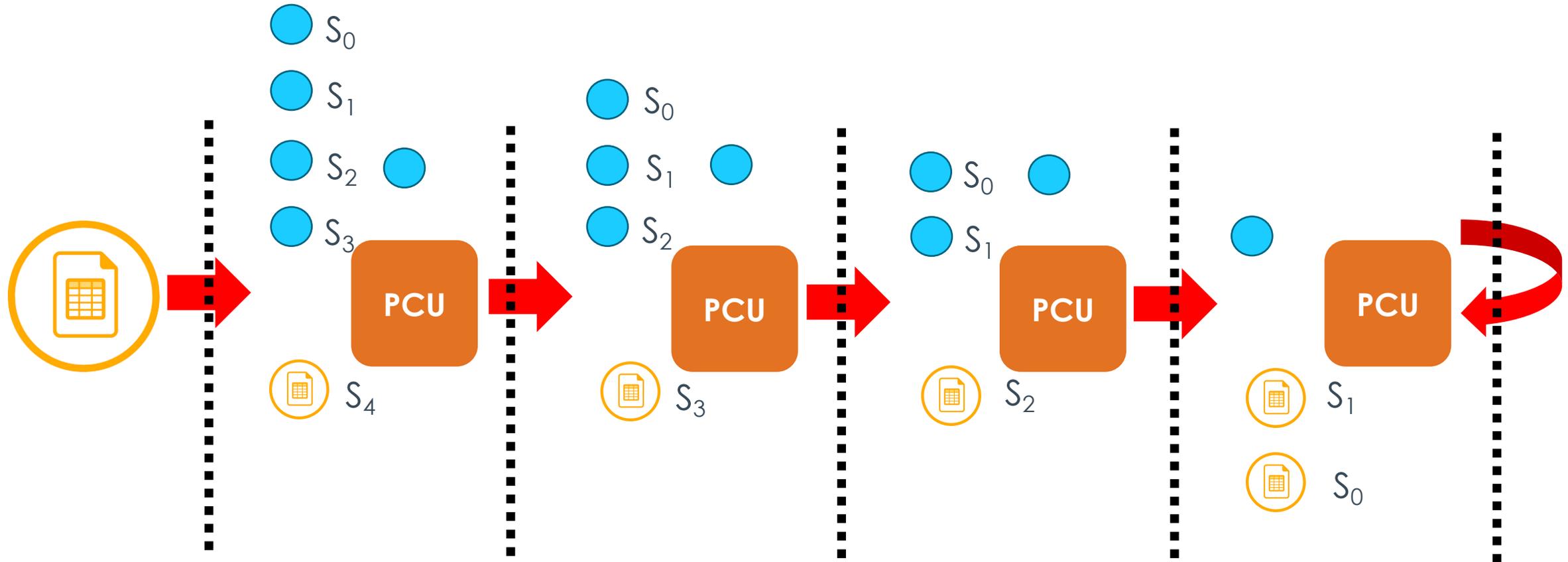
# HW Cost: PipeDream



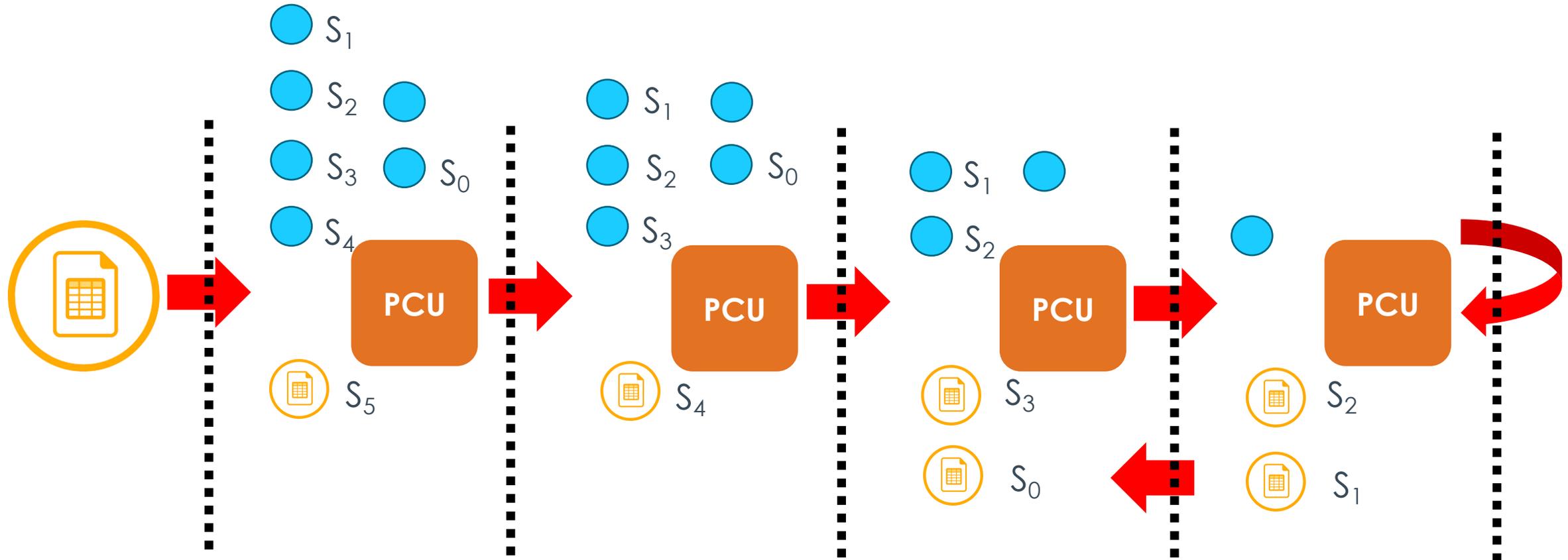
# HW Cost: PipeDream



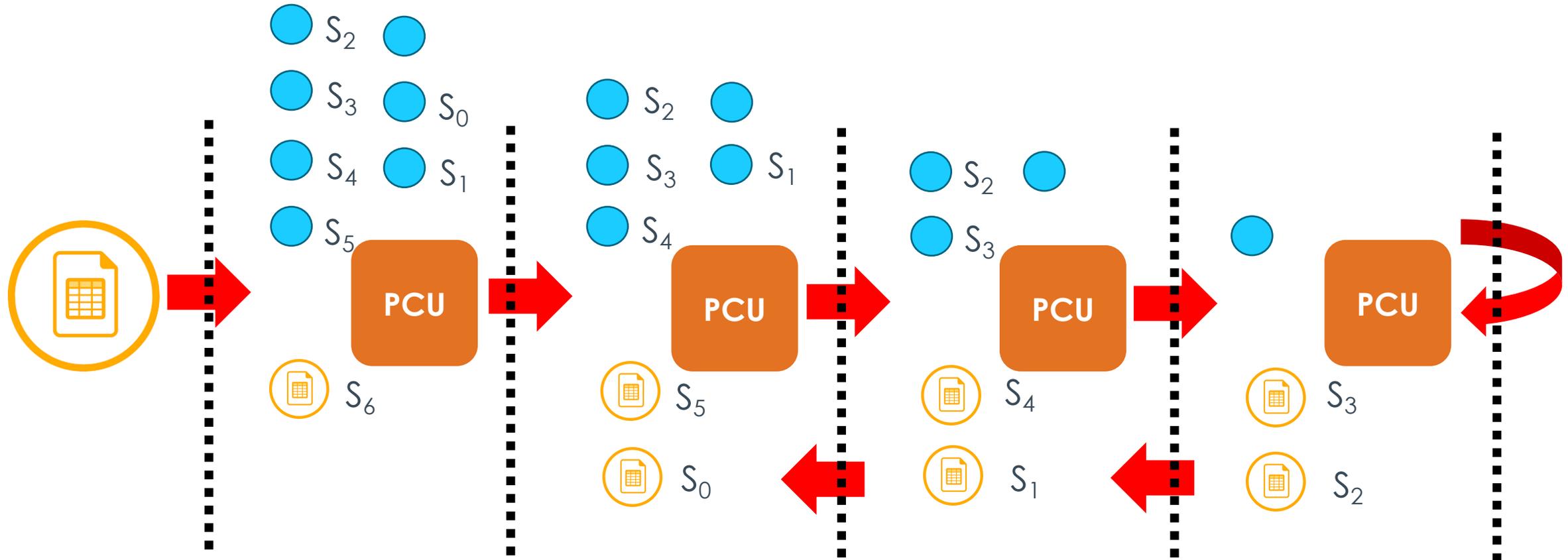
# HW Cost: PipeDream



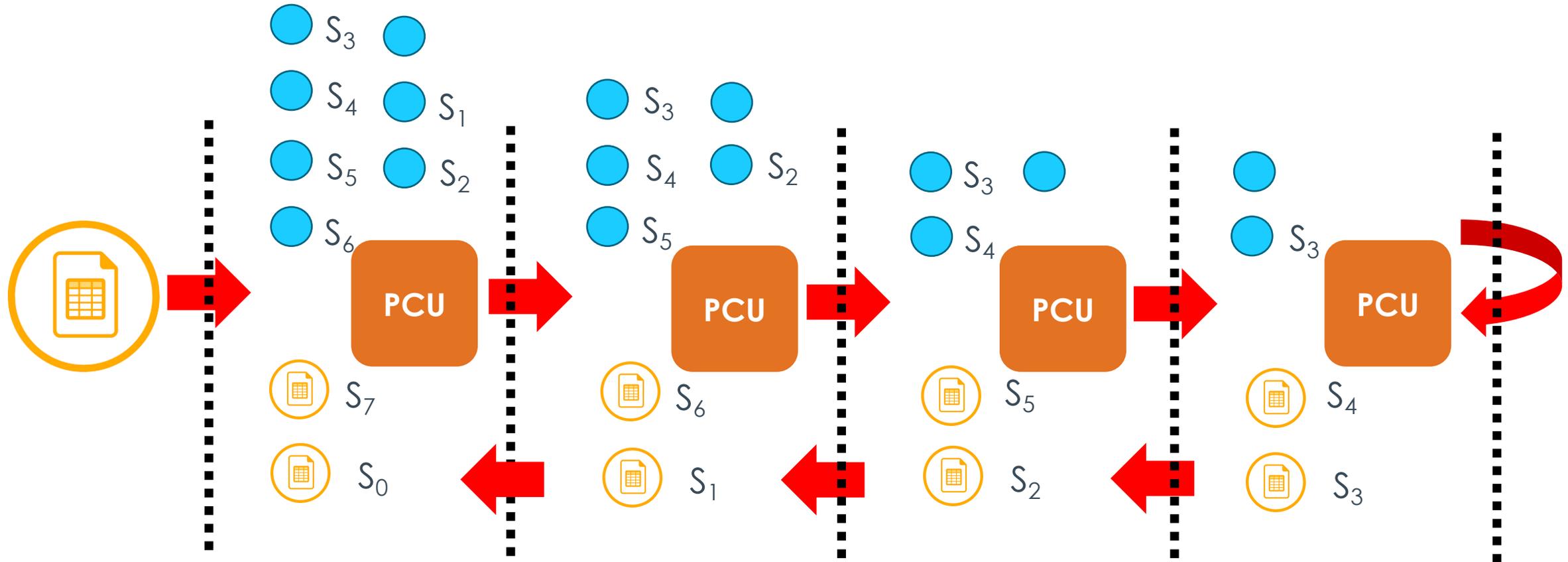
# HW Cost: PipeDream



# HW Cost: PipeDream

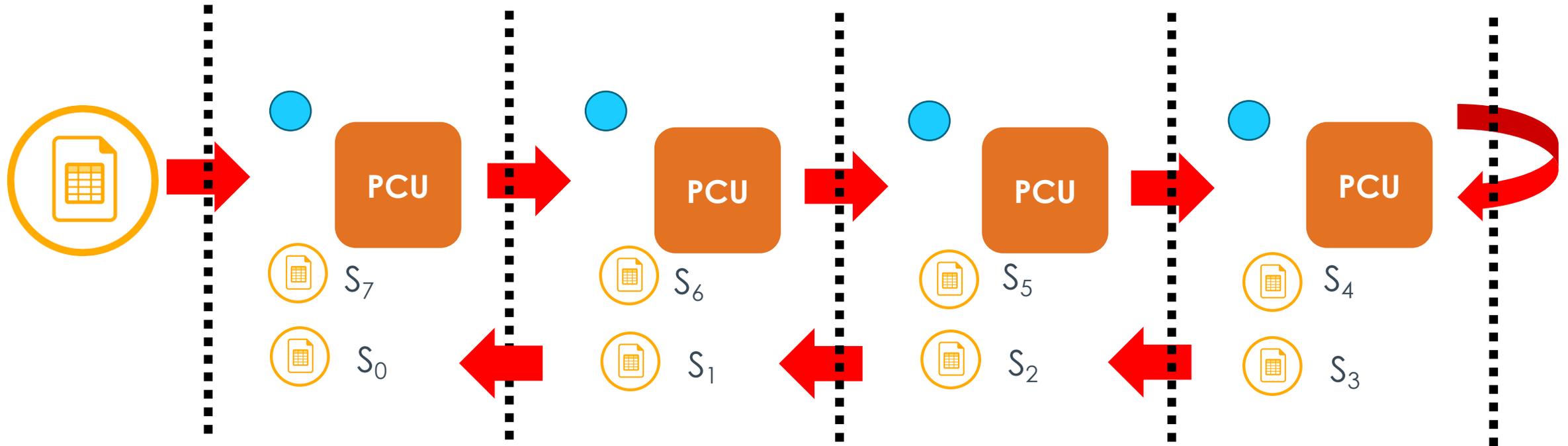


# HW Cost: PipeDream



**Panic:** Sacrifices memory for synchronous execution!

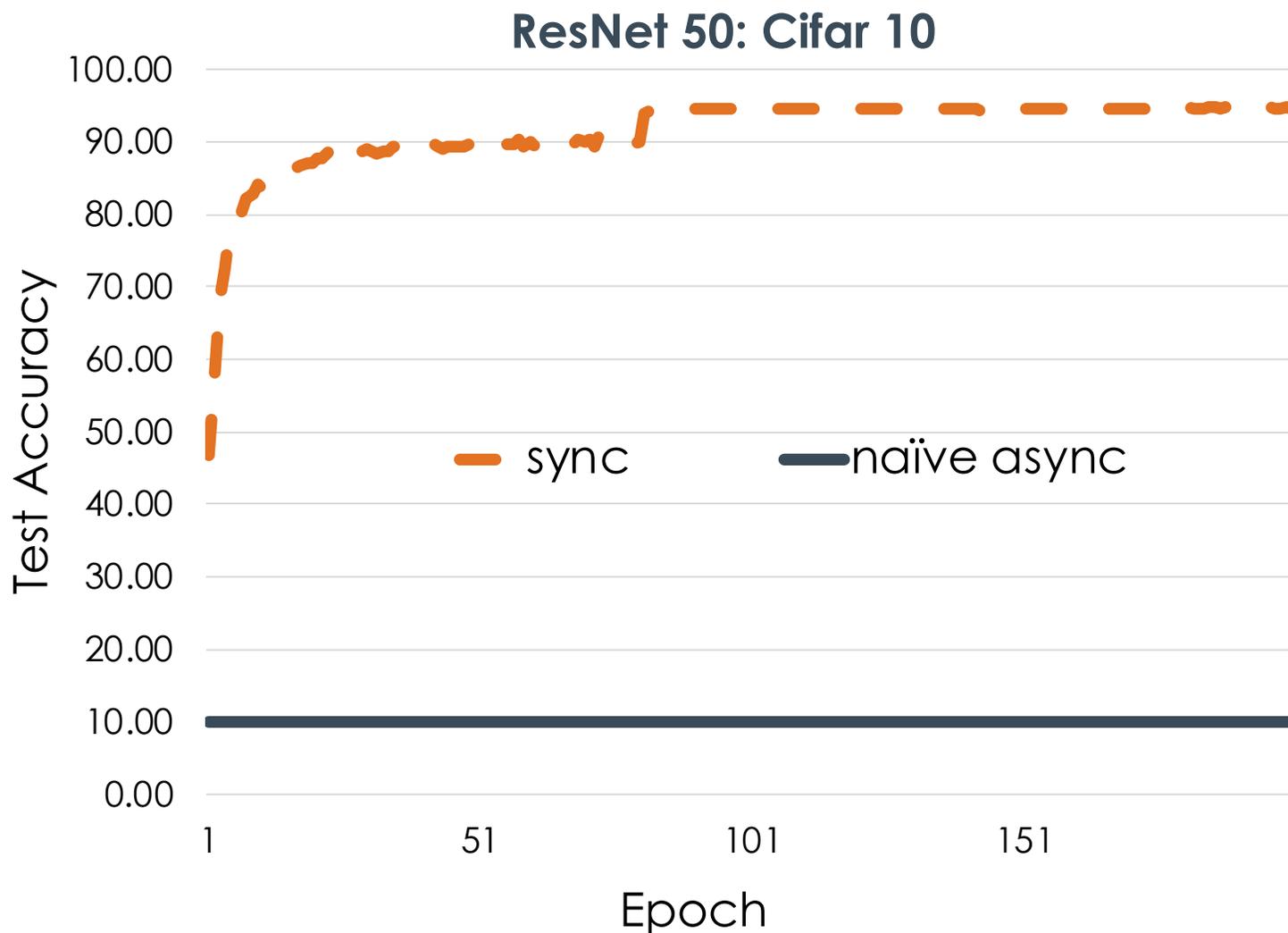
# Ideal Pipeline Parallelism Steady State



**Goal:** No hardware sacrifices!

**Panic:** Introduces **asynchrony** (delays).

# Houston, we have a problem.



**Key Insight:** Scale your learning rate proportional to the delay.

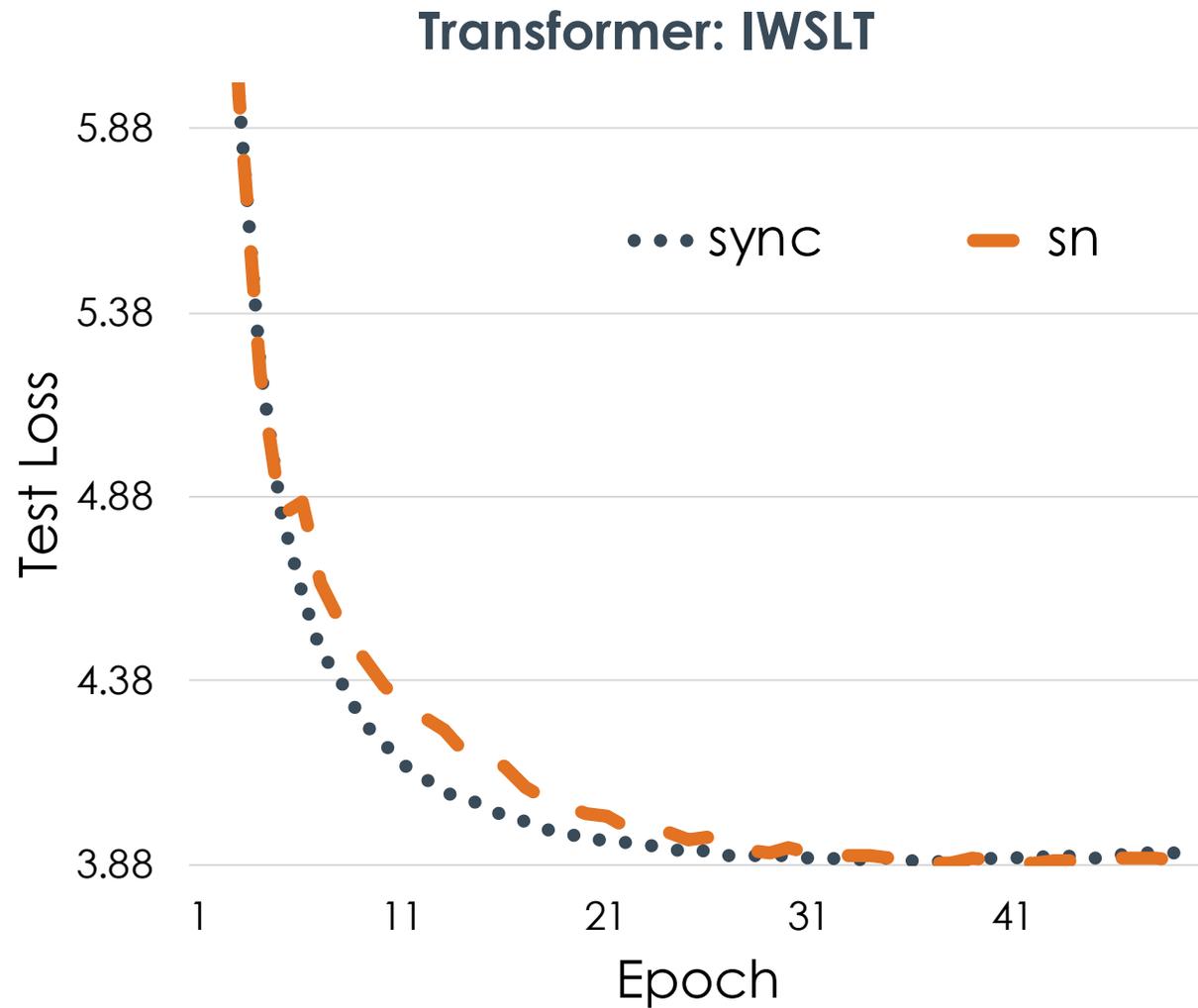
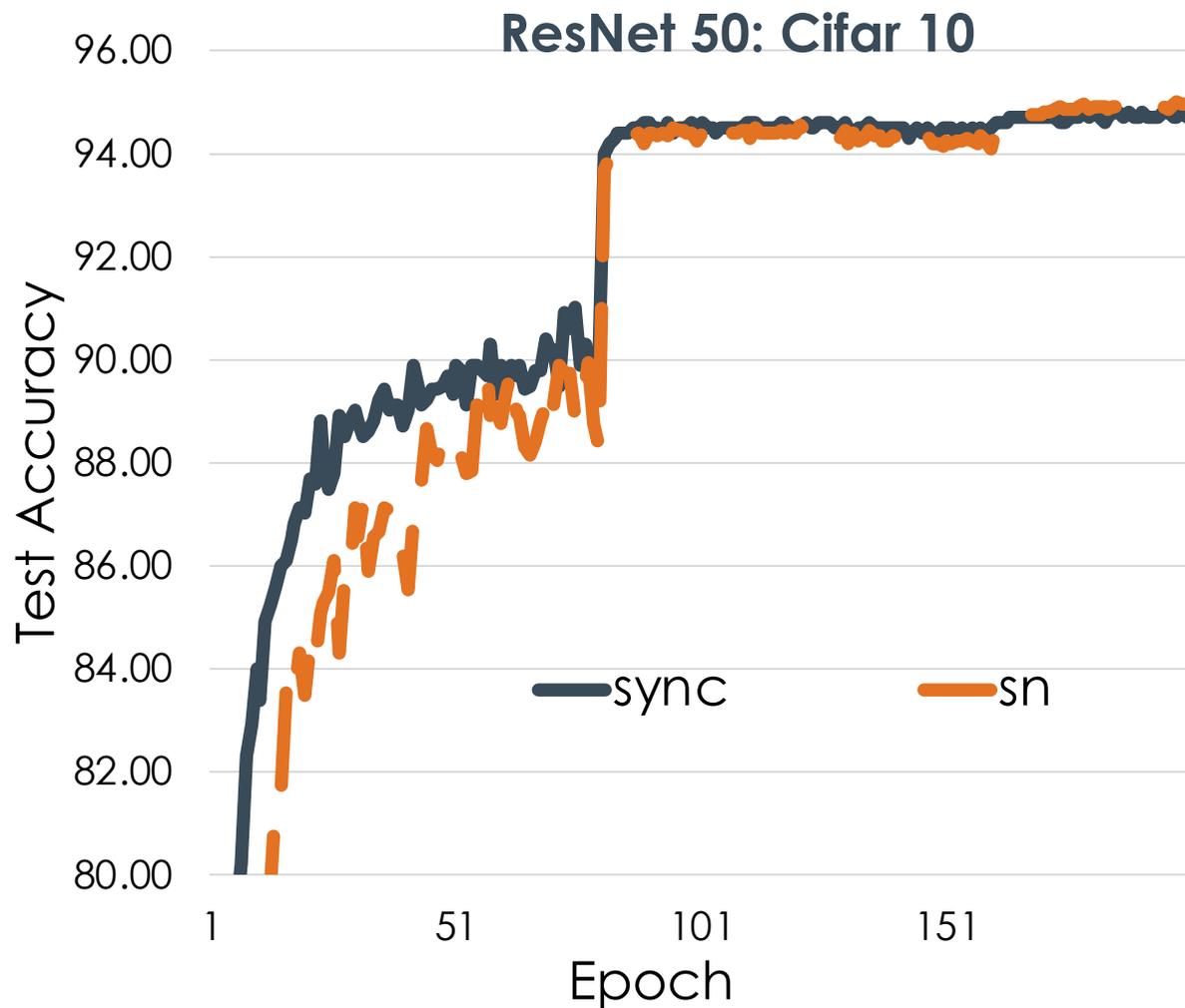
$$\alpha = \min \left( \alpha_{\text{sync}}, \frac{C}{\tau_i} \right)$$

Chris De Sa



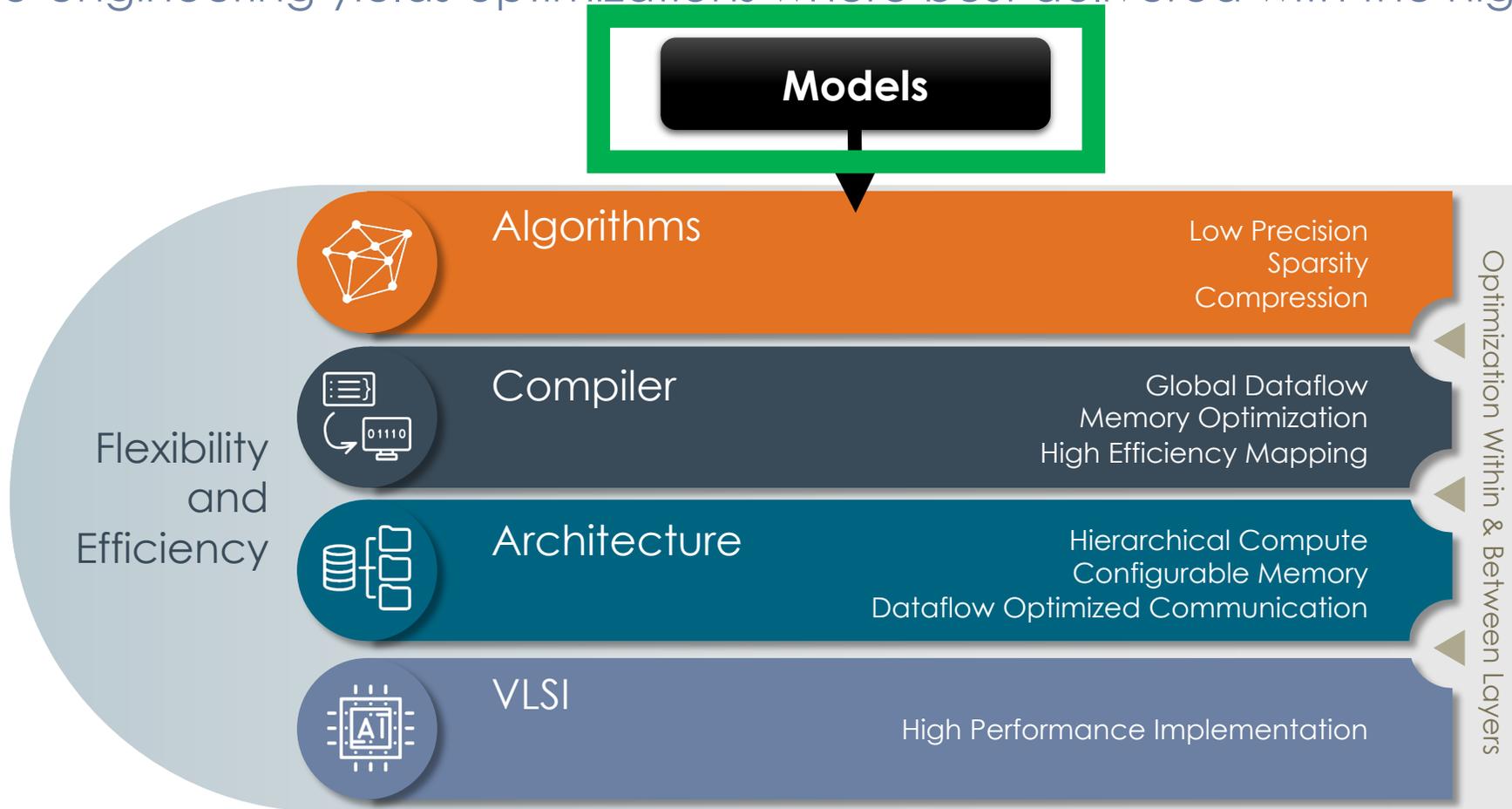
# Enabling Peak Dataflow Efficiency

**PipeMare: Arxiv '20**



# The SambaNova Systems Advantage: Reconfigurable Dataflow Architecture

Full stack co-engineering yields optimizations where best delivered with the highest impact

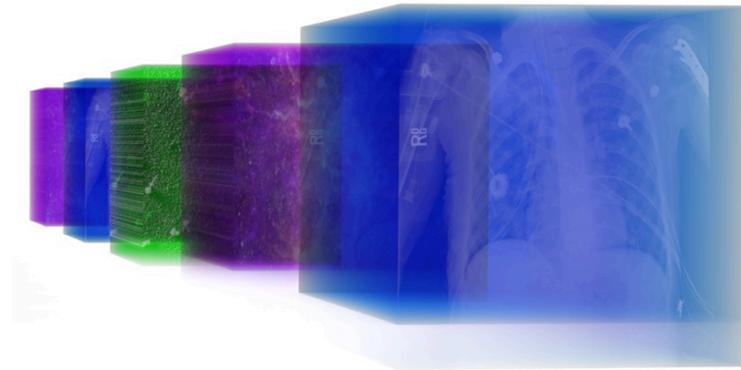


How do we future proof our code?

**What are the future models?**

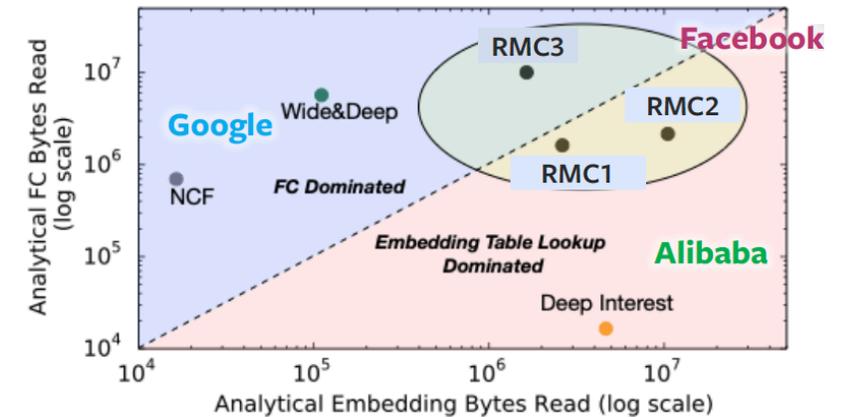
***Models are the  
new code.***

# Enabling New Capabilities (0 $\Rightarrow$ 1)



**Trillion parameter NLP models**  
Key to knowledge understanding

**High Resolution Deep Learning  
50k x 50k**  
Astronomy, medical imaging,  
X-ray imaging, ...



**Recommendation models with  
huge 100GB embedding tables**  
Recommendation is the  
backbone of internet services

# Part 1: NLP

*Models are the new code.*

# Proliferation of NLP Models



CreateML



FastBERT



Elmo, RoBERTa



BERT, XLNet



Microsoft

MT-DNN  
Zero



OpenAI

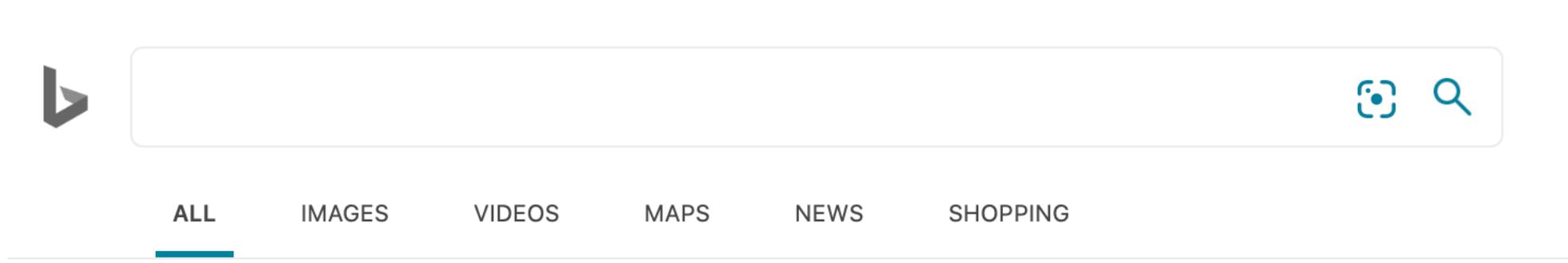
GPT, GPT2



salesforce

CTRL

# Richer Context, In a Small Amount of Space



Microsoft open sources breakthrough optimizations for transformer inference on GPU and CPU

January 21, 2020



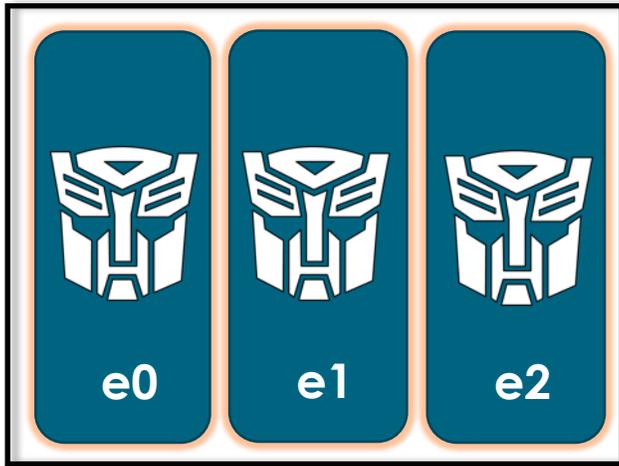
**EMMA NING**

Senior Program Manager, Azure Machine Learning

A **three-layer** BERT model in production at Bing.

Richer context, same space.

# Richer, Contextual Information



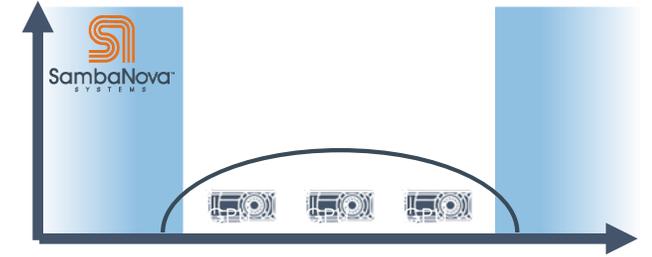
3-wide encoders

vs

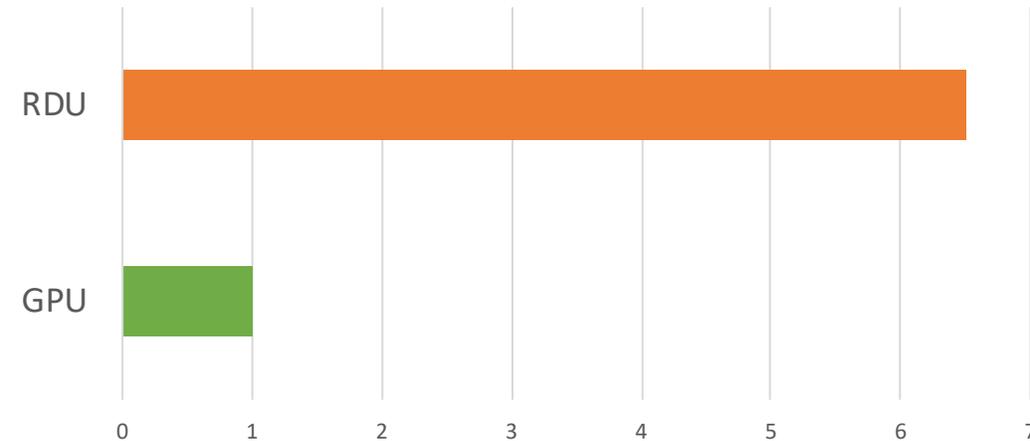


24-slim encoders

Fewer Parameters, Better Quality  
on **Natural Language Inference**  
QNLI : 3-layer 78.7 vs. Deeper 79

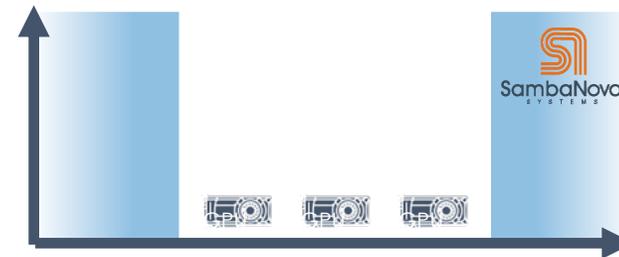


More than 6x faster on Deeper BERT



SambaNova enables Deeper Design Points

# Pushing the Boundaries of NLP

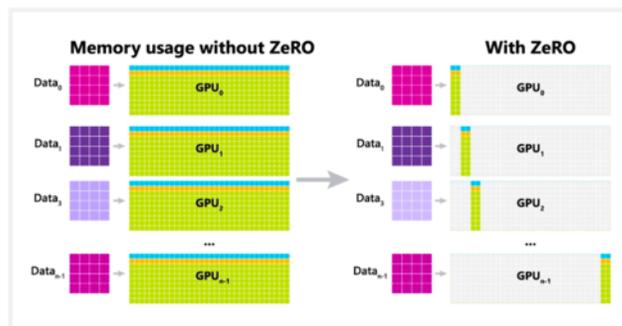


ZeRO & DeepSpeed: New system optimizations enable training models with over 100 billion parameters

February 13, 2020 | By DeepSpeed Team ; Rangan Majumder; Junhua Wang



## DeepSpeed + ZeRO



### Scale

- 100B parameter
- 10X bigger

### Speed

- Up to 5X faster

### Cost

- Up to 5X cheaper

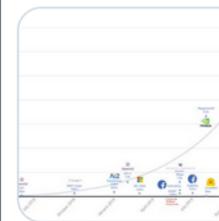
### Usability

- Minimal code change



**Satya Nadella** ✓  
@satyanadella

The new language model our teams built is the largest and most powerful one ever created – a milestone with the promise to transform how technology understands and assists us.



Turing-NLG: A 17-billion-parameter language model by Mic...  
This figure was adapted from a similar image published in DistilBERT. Turing Natural Language Generation (T-NLG) is ...  
[microsoft.com](https://microsoft.com)

9:28 AM · Feb 12, 2020 · [Twitter Web App](#)

# Enabling Large Model Architectures With a Single System

*Order of magnitude performance improvement, an order of magnitude fewer systems*



64 DGX-2  
1,024 V100s  
32 TB HBM  
16 racks



8 RDU,  
12 TB DRAM,  
1/4 rack

1 DataScale system

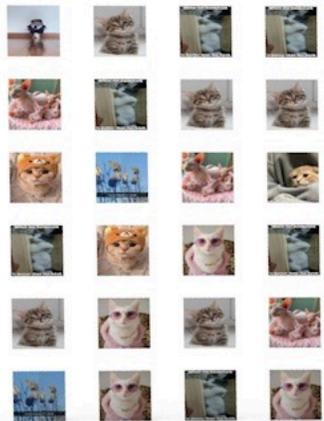
“One Model” 1 Trillion Params in a Single System: **Same** Programming Model

# Part 2: Vision

*Models are the new code.*

# Fast Growing Scale of Model Training Data

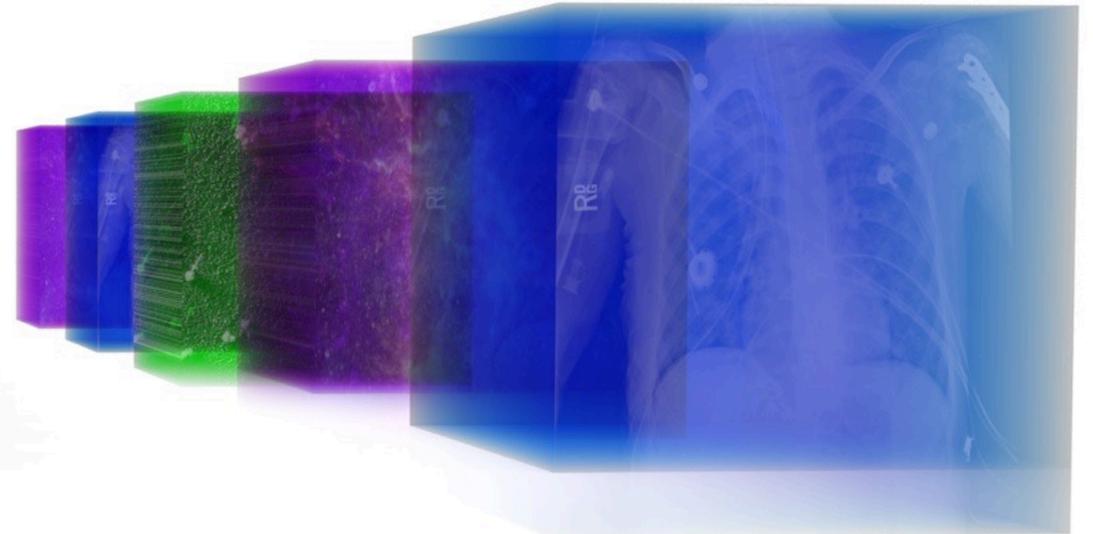
## Evolution of high-resolution Deep Learning



**Low-resolution**  
(e.g. cats)



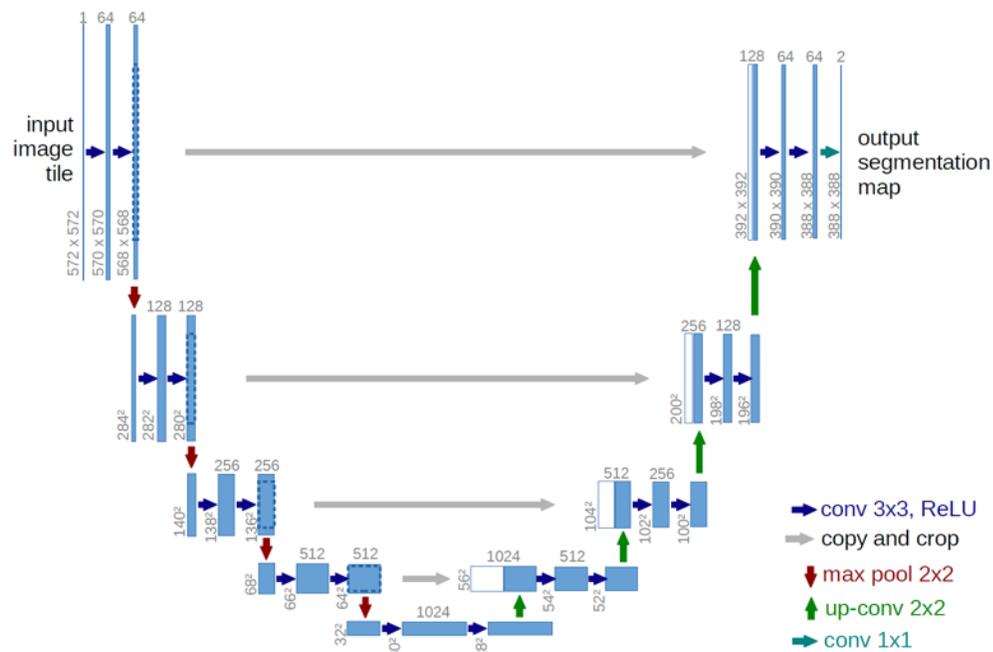
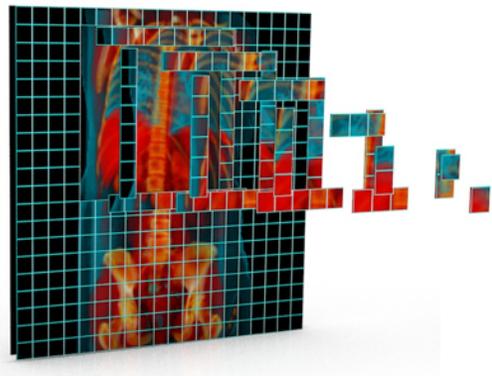
**4k images**  
(e.g. Autonomous driving)



**50k x 50k**  
(e.g. astronomy,  
medical imaging, virus, ...)

# Mapping High-Res Images to SambaNova

40k x 40k image running forward pass on UNet (image segmentation model)



**Tiles are streamed through model pipeline on chip**

- 3 x 40960 x 40960 input
- 409600 tiles per surface, or up to 26 million tiles for 64 channels
- GPU fails to allocate memory
- Even CPU errors out in PyTorch!

**RuntimeError:  
offset is too big**

**Only SambaNova can run these workloads out-of-the-box**

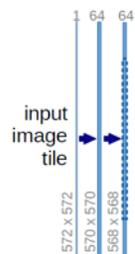
# No Compromise High-Res

**Classic tiling:**  
chop image  
into sub-  
images

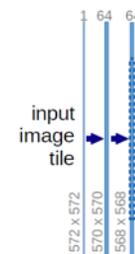
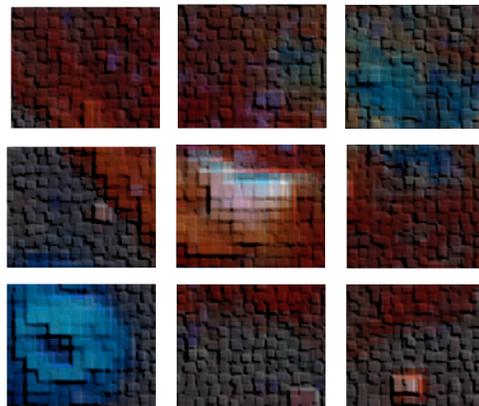
**Loses  
information  
in output!**



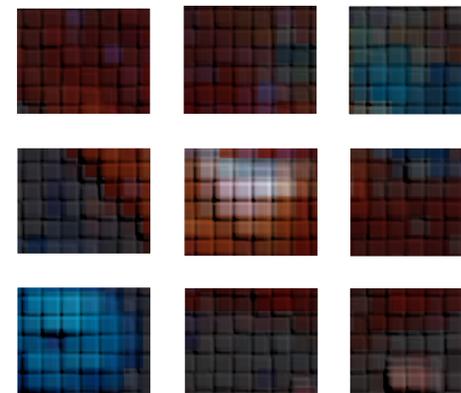
Tiled input



conv



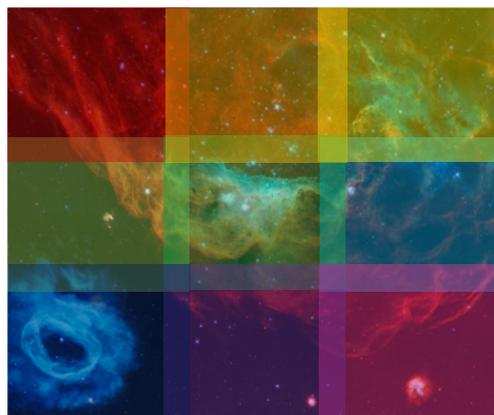
conv



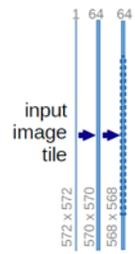
Tiled output

**SN tiling:**  
handles  
overlaps  
across tiles  
based on  
network

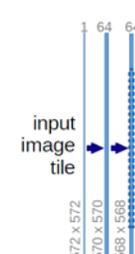
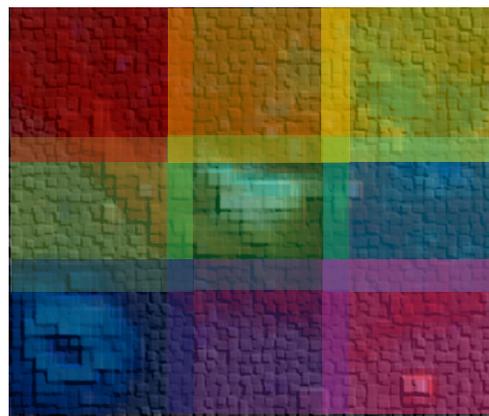
**Identical  
result as  
non-tiled!**



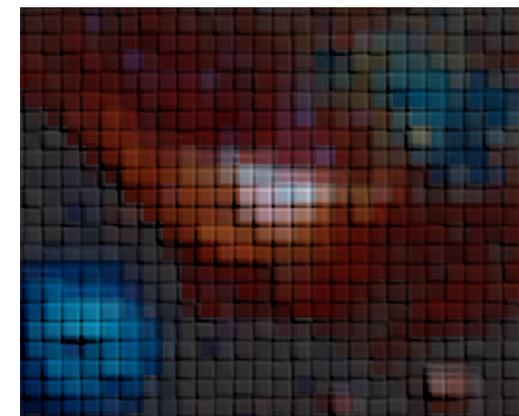
Tiled input



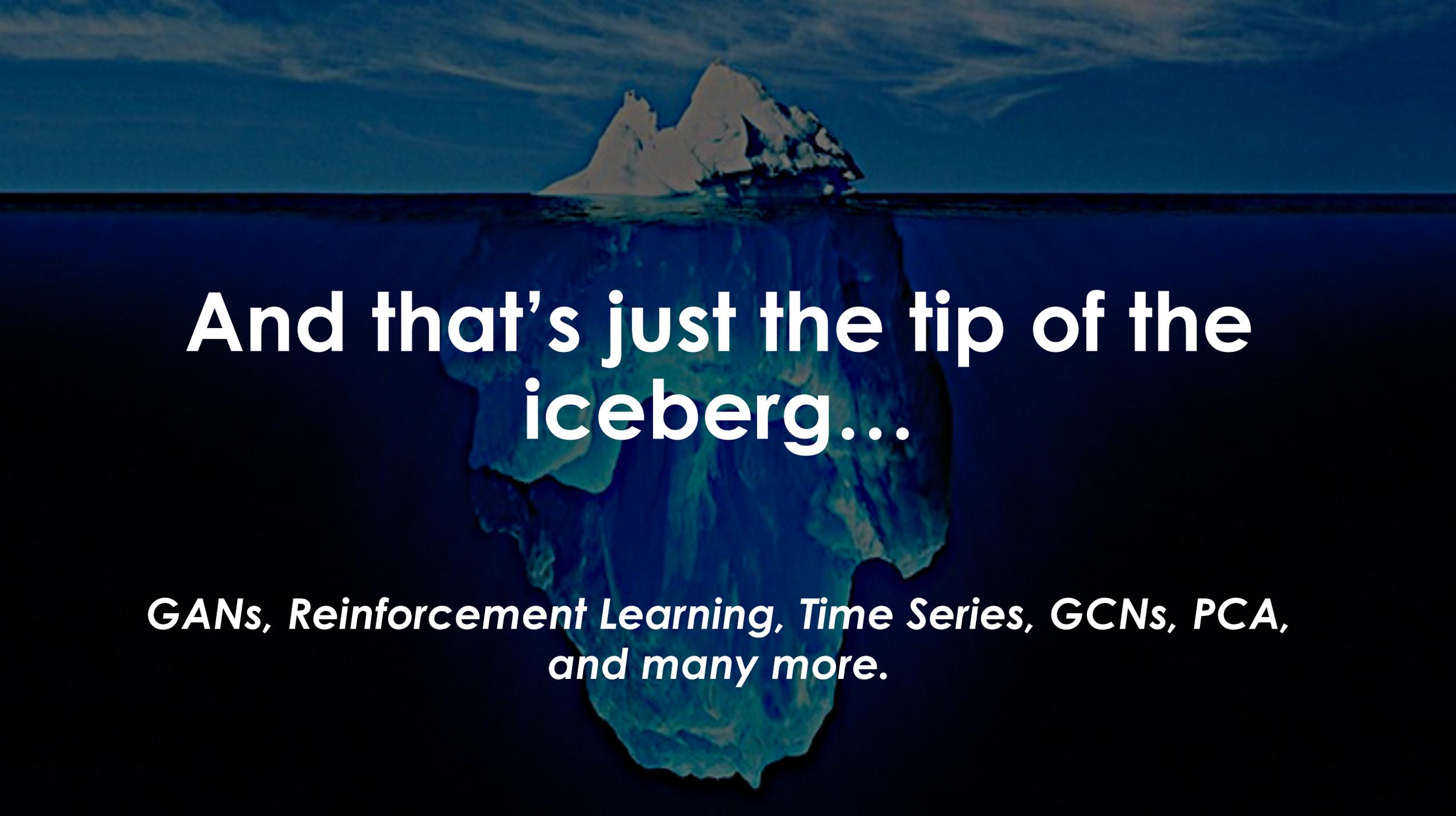
conv



conv



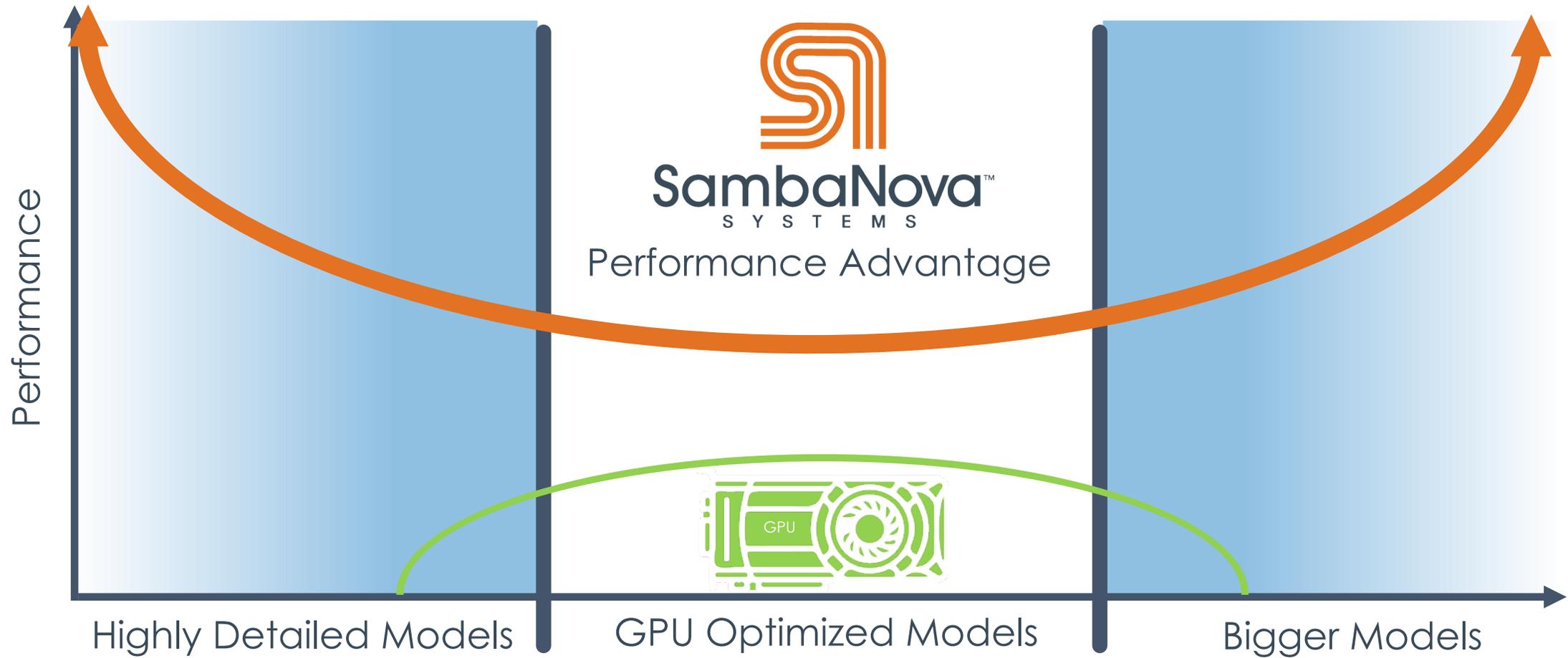
Full output

An iceberg floating in the ocean. The tip of the iceberg is visible above the water line, while the much larger, submerged part is visible below. The sky is a deep blue with some light clouds, and the water is a darker blue. The overall tone is somber and metaphorical.

**And that's just the tip of the  
iceberg...**

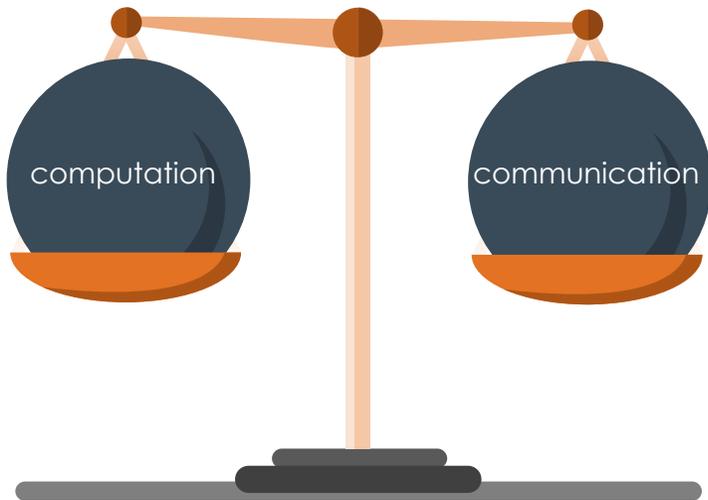
***GANs, Reinforcement Learning, Time Series, GCNs, PCA,  
and many more.***

# SambaNova: Breaking the Goldilocks Barriers, for Everyone

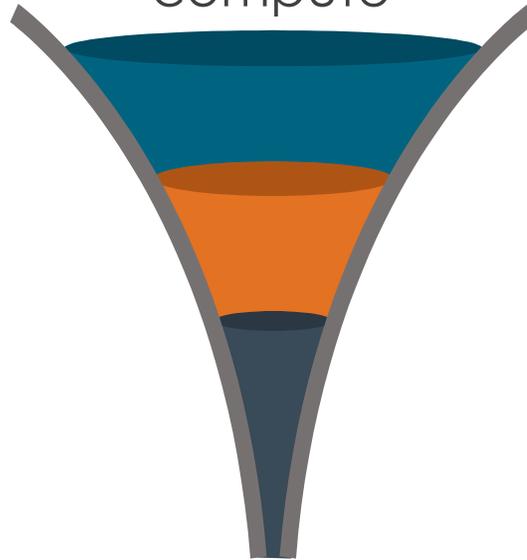


# Reconfigurable Dataflow for Unprecedented Flexibility

Performance  
balances  
computation &  
communication



**Bottleneck:**  
Yesterday's platforms  
only program  
compute



Flexibility unlocks:

- 10x **performance**
- 0-to-1 **applications**



We're hiring: [sambanova.ai](https://sambanova.ai)